INSTITUTE OF PUBLIC HEALTH

COLLEGE OF MEDICINE AND HEALTH SCIENCE

UNIVERSITY OF GONDAR

Application of data mining in prediction of anti-retroviral therapy outcomes among HIV/AIDS patients of Adama referral Hospital, Ethiopia, 2012 G.C.

Student: **Nebiyou Wendwessen(BSc)**

Advisors: **1. Dr. Berihun Megebiaw(MD,MPH)**

2. Mr. Bikes Destaw (BSc, MPH)

A THESIS SUBMITTED TO THE INSTITUTE OF PUBLIC HEALTH, COLLEGE OF MEDICINE AND HEALTH SCIENCES, UNIVERSITY OF GONDAR IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF PUBLIC HEALTH

JUNE 2012

GONDAR, ETHIOPIA

INSTITUTE OF PUBLIC HEALTH

COLLEGE OF MEDICINE AND HEALTH SCIENCE

UNIVERSITY OF GONDAR

Application of data mining in prediction of anti-retroviral therapy outcomes among HIV/AIDS patients of Adama referral Hospital, Ethiopia, 2012 G.C.

By: Nebiyou Wendwessen (BSc)

E-mail: nerwtmk@yahoo.com

Approved by examining board

_____ _____

Head, INSTITUTE of Public Health

Advisors

1. **Dr. Berihun Megebiaw(MD,MPH)** _____
2. **Mr. Bikes Destaw (BSc, MPH)** _____

_____ _____

Examiner

Acknowledgement

I would like to forward my deepest gratitude to my advisors Dr. Berihun Megeibaw (MD, MPH) and Mr. Bikes Destaw (BSc, MPH.) for their meticulous advice and valuable comments in all stages of the preparation of these thesis report with their full compassion, interest, encouragement and constructive criticism.

I would also like to forward my heartfelt thanks to all my family, especially to my mom Tejinesh mira, who helped me with all she could, my father, younger brother and sister.

I would like to acknowledge staffs of the ICT, classmates and colleague for their material and academic support necessary for my thesis. I also would like to thank staffs Adama referral Hospital for their cooperation and permission to conduct the study and University of OSLO, which is supporting the Health Informatics program at University of Gondar.

Above all, I thank God who helped me in all.

Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ANNEXES

ACRONYMS

AIDS	Acquired Immunodeficiency Disease
AI	Artificial Intelligence
ARFF	Attribute Relation File Format
ART	Antiretroviral Therapy
AUR	Area Under ROC
CRISP-DM	Cross-Industry Standard Process for Data Mining
HIV	Human Immunodeficiency Virus
PEPFAR	President's Emergency Plan for AIDS Relief
PLWHA	People Living with HIV/AIDS
RF	Random forest
ROC	Receiver operating characteristic
SVM	Support vector machine
WHO	World Health Organization

Abstract

Introduction: Data mining is the process of finding interesting hidden knowledge in large database. Since, antiretroviral therapy service in Ethiopia started in 2003 there is large amount of data gathered and stored in large databases. Due to lack of appropriate data analysis technique, this data were not used to overcome early detection and prevention of unintended outcomes from antiretroviral therapy.

Objective: the aim of this study is to predict outcomes of antiretroviral therapy among HIV/AIDS clients at Adama referral Hospital, using data mining techniques.

Methods: Institution based retrospective follow up study in using Cross-Industry Standard Process for Data Mining study on seven years records of Adama referral Hospital antiretroviral therapy clinic were conducted from April 17 to 24, 2012. In this study, electronic database composed of 8 relational table and 19088 records for both ART and pre ART clients were taken. The database suffers from multiple missing value and outliers, therefore cleaned by substituting mean for numeric and mode for nominal values. For important attributes, complete deletion was utilized. Finally, 10,690 records of adult antiretroviral therapy ever started clients were taken. The selected dataset was transformed into ARFF. Moreover, model building and evaluation was applied in support of Waikato Environment of Knowledge Analysis (WEKA) version 6.6.6 machine learning software.

Result: Among reviewed 10,690 records of ART clients 26.0% were drop out from ART services. Accordingly, twelve experiment were conducted in using both J48 and multilayer perceptron algorithms. Among the experiments, decision tree from J48 algorithm with balanced data set showed 93% accuracy of prediction. In addition, adherence, month on ART and family planning utilization were the top three-predictor attribute selected in the model.

Conclusion and recommendation: in this study, prediction of ART outcome in using data mining techniques, showed the applicability of data mining for early prevention of dropout and lost follow up from the ART. For this reason, this data mining approach could be integrated in the Adama Hospital ART clinic database for future prediction of new clients' outcome.

1. Introduction

1.1. Statement of the problem

“Data mining methods, applied within the process of knowledge discovery in databases, enable discovery of complex patterns and used to predict human behavior in large data sets. This capability has been demonstrated in multiple studies of health-related phenomena”(1).

According to a 2011 reports of UNAIDS around 33.2 million people in our globe get infected and lives with HIV, From this number sub Saharan Africa accounts around 68% (22.9million) of all HIV-positive people(2). By the year 2010, in sub sharan Africa 1.9 million people were newly infected by HIV virus(3). Ethiopia has one of the largest prevalence of HIV/AIDS in sub-Saharan Africa, For the past five years, the HIV/AIDS prevalence was 2.2% and was expected to rise to 2.4% in 2010. There are substantial prevalence differences between urban (7.7%) and rural (0.9%) settings(4,5). By getting understanding of non-affordability fee based ART service for PLHIV in Ethiopia free ART service provision was started in January, 2005 through the fund provided by PEPFAR (US President’s Emergency Plan for AIDS Relief(6) and became available in 22 Hospitals, after starting of fee based ART in 2003 on 12 government hospitals. From 2004 to March 2010, 251,060 HIV-positive people in Ethiopia were ever started ART(7–10).

But this is a lifelong therapy in order to attain its objectives consistence and sustainable good adherence (90-95%) of PLHIV to the treatments is mandatory(11). Otherwise, clients may lost or dropout from the ART. In addition, clients may dead due to treatment failure. The cause of dropout from ART may related with lack of knowledge, loss of interest in ART ,use of herbal medication, forget fullness, lack of food, transportation problems, taking hard drugs; drinking alcohol, being bedridden; mental illness, holly water, imprisoned, living outside service area and having an HIV discordant partner (or unknown HIV status) (9,12–15).

According to WHO in Ethiopia approximately 15% of all patients drop out from chronic care of HIV/AIDS Other Study conducted in Jimma university in 2008, show that, of 1270 patients who started ART, 173 (13.6%) were drop out from ART, 355 (28.0%) were belongs to lost follow up and 75 (5.9%) were died(9). This figure showed that high existence of un intended outcomes of ART dropout, lost follow up and early death. Dropout and lost follow up from ART has negative impact in attaining maximum sustainable viral load supersession and lead to high AIDS morbidity, hospitalization and mortality(2,14,16). Moreover, dropout from ART is one of the major causes of drug resistance and treatments failure(9,17). Another retrospective study in 2010 in Ethiopia 37% of AIDS patients on ART Commit risky sexual intercourse(12). This figure show as there will be high transmission of even drug resistant viral strains. This has been catastrophic event for the society in accordance with high dropout rate from ART.

The aim of this study is therefore to predict the outcome of ART among new clients using data mining classification technique based on the existing client's records. In addition, the study was discover hidden knowledge for high PLWHA retention in ART service.

1.2. Literature review

1.2.1. Overview of Data Mining

The advantage of Modern machine learning (ML) techniques over traditional statistics in future prediction and estimation tested in different application now a day's. Some of its application has significant acceptance like, in finding interesting information on predictive toxicology, disease classification, selective integration of multiple biological databases, functional neuro-imaging, etc(18–20).

Data mining is a branch of computer science(21) it's main aim is to uncover the hidden knowledge within the data sets and to predict outcomes the feature value of the variables based on fitting model(22) and extract important information from large data base(20,23). On other hand, data mining can shows a series of patterns that existed in large data sets(21,24). "The general algorithm for data mining consists of three parts.

1. The model

2. The preference criterion (a model set of parameters)

3. The search algorithm"(22).

Based on their function Data mining methods can be classified as exploratory data analysis, descriptive, predictive and pattern discovery. Predictive data mining is a kind of supervised learning and the value of the response variable where known in advance. Models like classification and regression are included in this category. Whereas, clustering (segmentation) algorithms, pattern recognition models, visualization methods are belong to descriptive data mining. Descriptive data mining is a kind of unsupervised learning because there is no already-known result to guide the Algorithms(22,24).

There are three common tasks of data mining: clustering, classification, and association rule learning(25).

1.2.1.1 Classification

Classification is a method that can build a model to distinguish and classifies data's in to classes, It also facilitates predicting the class label of the patients or their status of health which is not known in advance by using the existing models that are constructed by a number of classification algorithms(20,26). classification is a two-step process, model construction and model using for classification(24).

From classification algorithms, Due to the reasons that it's easy and simple tool for understanding and interpreting the results decision tree has got popularity in medical data mining(26,27).

A. Decision tree (J48)

A decision tree is a powerful analytical tools that produce interpretable results and constructed in combination of multiple nods and leaf,(28) each interior nodes associated the inputs of each variables and each leaf of the decision tree corresponding to the value of the given variable. The weakness of J48 lies with its capability in generalization of building the model(21,29). Decision tree strength lies on its capability to handle categorical data or non-numeric data and minimizes the amount of data transformations and the explosion of predictor variables(24). Decision trees have been found useful In HIV research, it has been used to analyze the association of antiretroviral resistance mutations with response to therapy and to predict drug resistance based on HIV mutations(26,29). The most popular algorithm for inducing decision trees is C4.5 (J48), an extension of ID3, that was developed by Ross Quinlan(26). The J48 decision tree based on C4.5 algorithm divides data in to group of data sets by using iterative splitting process then it builds decision tree starting from root node by utilizing all training data sets. In this algorithm partitioning threshold calculated by selecting values that result in the largest decrease in impurity. A split will considered pure if after the split, all instances at the downstream nodes belong to a single class. Additional branches in the tree will added in a similar fashion until a specified purity threshold is reaches, and instances at each terminal node will assigned

a class label. Pruning options were varied here, for optimization(19). Final objectives of Decision trees mapping a series of rules that can predict lead to a class or value(24).

B. Multilayer perceptron network (MLP)

MLP is one of the most widely used neural network architecture .It can form a model for both regression as well as classification algorithms that maps non-linear problem. MLP also, provides a non-linear mapping structure from a real-valued input depending on the interpretation of the output(s)(22). A MLP network is also expressed in one or more hidden layers that works an approximator this approximation can be done with only one hidden layer or more for continuous functions(30). A MLP is a feed forward neural network process with non hidden layer is enough for two class classification by using activation function of one node layer(22,30). The main idea in MLPs is that the input vector is successively modified through multiplication by weight matrices in the different layers, and the products are transformed by non-linear activation functions(22,28).

1.2.1.2 Clustering

Clustering is an algorithm of finding finite number of clusters from the given data set that describes the data by magnifying there similarity within the cluster and dissimilarity between clusters . Unlike classification, the classes of the data not determine in advance before the data set pass through the processes of data mining procedure(34). During clustering data's are break down in to mutually exclusive set of clusters. Through clustering, we can visualize the density of distribution in the data set clusters and we can identify patterns and correlation between variables. Clustering technique can be applied in pattern recognition, data analysis, image processing, and market research. K-means algorithm is commonest algorithm in clustering. K-means algorithm reduces the cluster-sums-of-squares in order to assign the variable in to the given cluster(24).

1.2.1.3 Association rule analysis

Association rule analysis most widely used in business transaction and also, it has important application in observing patterns in biomedical data(29). Its application relay on actions to developing a model that gives as a rule of an association confidence to describe the occurrences of an events or records together. In other word, the probability that a transaction that contains X also contains Y and its support is the percentage of transactions that contain X and Y(34). Association rules combine events that are occur simultaneously to find interesting pattern. Association discovery used for to handle existing customer and in finding new customers, identifying potentially harmful behavior, for Web site navigation and Medical diagnosis/research(35).

1.2.2 Data Mining Applications in the Health Care

By its nature, medical records are huge and growing exponentially. To store and analysis this large medical data base automated and sophisticated algorithms may required(24). Data mining application architecture can handle this huge database for analysis and knowledge discovery(20).

Using data mining techniques to predict hospitalization of hemodialysis patients is explore in Taiwan. In this study decision tree (C4.5) is utilized for rule discovery. The investigation shows 99-100% accuracy for prediction(36).

Data mining also used to predict chronic illness that is occurs in the late age. Another study in Taiwan to predict cerebrovascular disease adopted three classification algorithms, decision tree, Bayesian classifier and back propagation Neural network. From the algorithms were used decision tree model has, better sensitivity and accuracy 99.48% and 99.59%, respectively(27). On other hand, mental health is one of the components of health service with huge amount of records. In Portugal, data mining were used to predict dementia. From the data mining algorithms SVM, logistic regression, RF and classification tree were used. In experimentation SVM showed that

76% accuracy and AUR 0.90, RF has also got 73% and 0.73 accuracy and AUR respectively(28).

In Ethiopia, there were some attempts to apply data mining in heath information system. Study conducted in on 2001 child labor survey database to understand the nature of child labor problem in Ethiopia. The investigator tries to identify relation between variables of the survey database. From the study, ten best rules are explored with minimum support of 90% and minimum confidence of 95% threshold(35).

Another study in Ethiopia conducted Butajira rural health project area. From data collected in Butajira rural health project area, to predict child mortality the investigator utilizes neural network and decision tree. The study found that an 93% accuracy in neural network and 95% accuracy decision tree with the default value of the parameters(24).

1.2.3 Application of data mining on HIV/AIDS Data set

A study conducted in south Africa, Johannesburg on adaptive control of HIV status of Individuals by using neural network application in order to understand how demographic variables like education level affect the risk of being HIV positive found that a prediction accuracy of 88%(37).

Another study conducted in India by applying MLP to classify HIV/AIDS infected and non-infected status of individuals The findings conclude that the MLP network algorithm produced the best performance with 89.80% accuracy(30). To predict the survival of AIDS patients in Malaysia by using fuzzy neural network prediction based on their CD4, CD8 and viral load counts the study found that100% accuracy based on the selected variables(38). In order to predicting HIV status of individuals based on Demographic and medical history information obtained from annual South African, Johannesburg antenatal surveys. The investigator use MLP then the investigator found that best performing MLP network has a training AUC of 0.7385 and a validation AUC of 0.6701, With an accuracy is 68% on training to estimate the risk of acquiring HIV(22).

1.2.4 Data mining application on antiretroviral therapy data sets

As the knowledge of the principal investigator concerned, there is no attempt to apply data mining to predict outcomes of ART in Ethiopia. However, there are studies related to ART database in Africa, western countries and Asia. Some of them are:-

Data mining application can predict the response of patients for specific treatments. Patient's response to ART also predicted by study conducted in Rome, Italy. The investigators use a combination of expert rules, logistic regression and non-linear machine learning. In this investigation logistic regression found that AUC 0.76 and accuracy of 75.63% and RF found that AUC of 0.77 and accuracy of 76.16%(39).

Study conducted in Stanford university on HIV mutation changes based on treatment history investigators uses combined data mining approach to predict infrequently occurring HIV mutation given that history of ART, from those approaches classification tree performs 56-67% 0f AUC measurements with 0.46-0.66 sensitivity and 0.53-0.76 specificity predicting performance in each coding position(29).

Another study conducted by using Random Forest (RF) classification data mining method, to optimize antiretroviral therapy without genotype resistance testing. Study is done based on retrospective data from European merged cohorts study, found that with an average AUC 0.77 vs. 0.757 at 8-weeks treatment change episodes, 0.834 vs. 0.821 at 24-weeks treatment change episodes for both RF i and RF ii tests(40).

Study conducted in south Africa on prediction of CD4 count change using support vector machine classification model has found an accuracy of 83% prediction by taking genome, current viral load and number of weeks from baseline CD4 count as an input from publically available Datasets of Stanford HIV drug resistance database(41).

Having the knowledge that no single attempt to predict outcomes of ART in Adama referral Hospital and Oromiya region as whole in using data mining technique, this research project was done with aim of finding interesting predictive rule to prevent lost follow up, drop out and early death among ART clients.

1.3. **Justification of the Study**

ART improves immunological response of PLHIV to prevent AIDS related disease. Due to this the survival of ART users is increase by three fold than those of non ART users (7, 16). Despite its importance, for better quality of life and reduction of AIDS morbidity and mortality, poor adherence to treatment and lost follow up as well dropout from treatment are now becoming a major threats in ART programs(2,9,15).

There is high existence of lost follow up and dropout rate from lifelong treatments of HIV/AIDS at a national level in general and at Adama referral Hospital in particular. To alleviate these problems experts suggested that strong and ongoing adherence counseling is mandatory. However, achieving more than 80% adherence for treating chronic illnesses without having significant amount of dropout from treatments has been problematic(16). This challenging issue is stranger in resource-constrained settings where the health care services are not well developed. Because PLWHA has social, economical, psychological and medical factors, that affects their sustainable retention in the ART services at the time of their first contact and in the whole courses of the treatments. However, despite the importance of predicting those factors for preventing lost follow up and dropout from ART service, no attempt was made to find out a model that can achieve predicting outcomes of ART in advance until now in Ethiopian context based on those factors listed above, as to the knowledge of the principal investigator.

This study provides a model that will predict outcomes of ART among new clients. In addition, this study using data mining technique will help Health-care professionals, planners and policy makers by showing important direction to prevent lost follow up, dropout from ART and early death through the discovered hidden knowledge.

2. Objective of the study

2.1. General Objective

- The general objective of the study is to apply data mining techniques in predicting outcomes of lifelong antiretroviral therapy among HIV/AIDS patients of Adama referral Hospital, 2012 G.C.

2.2. Specific objectives

- To identify attributes for predicting outcomes of ART.
- To prepare the data set for model building.
- To identify suitable algorithm for model building.
- To assess the model performances in predicting outcomes of ART.

3. Methods and materials

3.1. Study Design and period

Institution based retrospective follow up study in using Cross-Industry Standard Process for Data Mining step by step process on seven years records of adult clients who ever started ART study were conducted from April 17 to 24,2012 G.C.

3.2. Study Area

The study was conducted in Adama referral Hospital, Oromiya region, East Showa, 99 km from the capital city Addis Abeba, from April 17 to 24, 2012. The catchment population of the Hospital estimated to be 5 million (14). The Hospital has been providing ART service since 2005. The clinic provide services by having 1 physician, 5 ART nurses, 3 data clerk and three peer educators. At the beginning of the provision of the service the data related ART services to clients was collected using manual method, then after with collaboration of ICAP-CU Ethiopia electronic data base for recording, storing and utilizing information's of clients were implemented in 2009. This database (DB) composed of eight relational tables that have records related to socio-demographic characteristics, past medical history, investigation, treatment given and follow-up data of ART clients. For back up and collecting other client medical history, the clinic still use manual record according to the new health management information system (HMIS) standards.

3.3. Data Source

The data source for this study was all adult records who were enrolled in HIV/AIDS care and support service in Adama referral hospital and stored in electronic ART database of Adama referral Hospital ART clinic.

3.4. Study records

The study records for this study were all-adult records who ever start ART in Adama referral Hospital and stored in electronic ART database of the ART clinic from February 2005 to the date of data collection.

Inclusion Criteria:- Adult PLHIV ever started on ART and registered in ART electronic data base of Adama referral Hospital at the time of data collection with complete values of an essential attribute that suggested by domain experts.

Exclusion Criteria:- Ever started on ART adult PLHIV and registered in electronic database but records value for an essential attributes not complete were excluded.

3.5. Sample Size and Sampling Procedures

Having the knowledge, taking all the records make the classification model more stable in data mining, all records of Adama Hospital ART clinic EDB were taken for model building in prediction of drop out from ART.

3.6. Variables of the Study

3.6.1 Dependant variable/ Class Attribute

Drop out from ART

3.6.2 Independent Variables

- Socio-demographic characteristics: - Sex, Age, Marital Status, Level of Education, Religion, Place of residence, Employment status, having family.
- Baseline medical characteristics: - WHO clinical staging, CD4 count, functional status, family planning utilization.
- Anthropometric measurements: - weight.
- Other Characteristics: - adherence, month on ART and ART regimen in the last visit.

3.7. **Operational Definitions**

Prediction is a process of forecasting future behavior or estimation future value by classifying records based on past events(42).

Lost - patients missing their appointment for drug pick up at least for one to three months

Dropout - patients missing their appointment for drug pick up for more than three consecutive months

Stop - patients who remained on chronic HIV care but discontinued ART due to medical reason.

Transferred out - Those patients who were transferred to other ART clinic.

Transferred in - Those patients who were transferred out from other ART clinic or health facility and accepted accordingly in the ART clinic of Adama Hospital.

3.8. **Data Collection procedure**

The primary source of data for this research was the Adama Hospital ART clinic EDB. Demographic as well as the current state of medical, anthropometrics and other characteristics of each member in the program records were collected from the electronic database. The database contains records PLHIV that can be identified by name, MRN and unique ART.

During data collection process, first full de-identified back up records from Adama Hospital ART clinic EDB were taken. Then after the records were evaluated critically for attributes, which are most relevant for prediction based on domain expert's suggestion. This domain expert group includes professionals that have background knowledge about ART.

3.9. Data collectors

For data collection, integration and for some part of preprocessing of the data one data manager from the Adama Hospital with BSc in IT and two data clerks were used.

3.10. Data quality control

In the first stage of data collection from the electronic database the principal investigator were give orientation for one day to the ART data manager and data clerks. The principal investigator also follows each steps of extraction of the de-identified records. During integration and rearrangement of the collected data the principal investigator were follow each process and again the preprocessed data were verified by regional ICAP-CU data manager for the integrated data consistency and accuracy of the attributes range value.

4. Data Processing and Analysis

For data, analysis and processing CRISP-DM step by step method set as a guide to achieve the objectives of the study.

4.1. Business understanding

In compressive HIV/AIDS care and support program of Adama referral Hospital, services were integrated from four components that includes :-psychological, social, legal and clinical ones. Each component is very important to deliver holistic Care and support services for PLHIV. On other hands, the program is the composition of services, which has strong association each other. Which, includes Voluntary Counseling and Testing (VCT), the diagnosis and treatment of Sexually Transmitted Diseases (STDs), Tuberculosis control program (TB), Opportunistic Infections treatments (OI), and the Prevention of Mother to Child Transmission (PMTCT)

In the first counter of the client to the ART clinic from any of the hospital services area or outside the Hospital as transfer in to the clinic, the data clerks take the first part to say welcome. Then the necessary information that includes Demographic information, HIV care and family status, previous ART summary and other Patient encounter information collected as per the guiding data collection standards.

The next step involves segregation of the client according to the presence of critical illnesses. If the client is not in the position for life threatening conditions pre-educators takes part for counseling and life experiences sharing. The counseling session continues by ART nurses and the ART intake form, which is specifically left over for nurses were filling by. For the evaluation of the general health, conditions and the proper WHO staging of the client physicians and health officer were take part. Soon after general physical examination and history taking about the client, undergo. In addition, laboratory examination for CD4 counts, base line laboratory investigation and supportive laboratory examination in order to diagnosis the presence of opportunistic infections under taken.

Finally, the clients come back with all his/her investigations result, the physician decides whether the client full fill the recommendations to start ART or not. Then the client appointed for consecutive treatments and counseling.

The whole processes of Adama Hospital ART clinic accomplished in using one physician, one health officer, five ART nurses, three data Clarks and three pre-educators.

4.1.1 Data Mining Tool Selection

Before selecting, the data mining tool suitable algorithms for developing the model were identified. Based on their multiple application into medical researches J48 and multilayer perceptron were selected. Specifically for J48 algorithm was considered for model building, for its simple and easily interpretable decision tree output.

Selecting appropriate data mining tool was conducted after the problem domain becomes understandable. To select the appropriate data mining tool in predicting outcomes of ART the principal investigator were uses some criteria's, like, the ability to perform the selected algorithms, familiarity and cost of the software. Therefore, due to the above criteria's the principal investigator selects "WEKA" software. This software developed at the University of Waikato in New Zealand. "WEKA" stands for the Waikato Environment of Knowledge Analysis(43).

4.2. **Data understanding**

After setting up the problem and a rough plan for its solution, the principal investigator proceeded with the central item in data mining process - data. There are several things to be learned the data before the actual application of data mining techniques. Records collected from Adama referral Hospital ART clinic electronically registered and handled in using Access database. In this database, eight relational tables were used. Among them, most importantly the registration table and follow up table considered for predicting dropout from ART according to domain expert suggestion. In the registration table demographic and identification, variables were incorporated and in the follow up table, the updated histories in the last day of the client follow up contact records were included. The EDB contains 19,087 both pre-ART and ART client records until the day of data collection. Among them adult on ART client records with the necessary information accounts around10,690 records. Taking the whole records will provide better model stability, so all adult on ART client's records on ART clinic EDB were taken for model building. In this regard, Rows represent records whereas columns represent attributes

4.2.1 Description of the data collected

After the initial data collection, the new data set created on MS Excel that contained the 52 columns and 190 rows from the selected socio demographic and follow up tables of MS Access EDB. The extracted data set composed of both ART and Pre-ART clients. The records belongs to ART clients were 13,136 in number and the remaining 5,952. The data set also suffers from missing value, outliers and lack of standardize range of values. Missing values of the data set accounts 27.3% of the total records, outliers accounts 7.2% of the records value and five attribute were lack standardize range of value for example occupation and address of the clients.

4.3. **Data preparation**

The main goal of this step is cleaning and removing discrepancies and inconsistencies for model building(20). The significance of this data mining step may not be understandable until we face its value in the analysis step data preparation process includes:- data cleaning, data selection, attribute or feature selection, transformation and aggregation, integration and formatting(44).

4.3.1 Data cleaning

In real world databases suffer with multiple missing values and outliers. The Adama Hospital database has its own mechanism to prevent outlier but for some attributes by putting legal value for example for age and sex attribute had legal value to prevent outliers. However, the database suffers from multiple missing value and outliers for some attributes. WEKA could not open a data file unless it is clean and in required format(45). Mean and mode substation were used to handle those missing values. Mean substitutions were done for numeric attribute like age and weight of the clients missing value cleaning. Mode substations also conducted for cleaning missing values of nominal attribute like sex marital status of the clients. In addition, for the records had missing values of very important predictor attributes, were excluded from the data set. For example, records had missing value for attributes adherence and WHO clinical staging (OAWHO) was excluded from the study.

Outliers and noise data from the data set were considered as missing value of the given record. Accordingly, outliers and noise data handled by substituting the mean and mode of the attribute values.

4.3.2 Targeted dataset Selection

The whole target dataset may not be taken for data mining task. Irrelevant or unneeded records and attributes were usually eliminated from the data mining database before starting the actual data mining function.

Records were selected based on inclusion and exclusion criteria's. From the available dataset records belongs to pre-ART clients excluded from the dataset. The reduced numbers of client records were accounts 10,690 ART records.

Attribute selection using dimension reduction methods were performed. Dimension reduction has the goal of reducing the number of inputs in the classification algorithms. It has significant importance in reducing over burden in computational procedures, over fitting and poor classification efficiency. Among Dimension reduction methods this study utilizes manual User-defined composites method(19,46). In User-defined composites method domain experts suggest important attributes and the weight of each attribute were measured. On the other hand, selection of data set process were done based on the data mining goals, inclusion and exclusion criteria's, quality and technical constraints such as limits on data volume or data types for both attributes and records (row and column)(47). The attributes from the ART EDB composed of 13 irrelevant attribute that describes about code of clients, whether the intake form completed and the referring health facility code. Those 13 irrelevant attribute were excluded from the dataset reason that not in line with data mining goals, discovering interesting knowledge. From the remaining 42 attributes 11 attributes were redundant in nature were aggregated into three attributes. Attributes that had more than 50% missing values were excluded from the data set according to domain expert's suggestion. Five attributes were belongs to missing value more than 50%, for example height of clients. Finally, predictor attributes 23 in number were given to the domain experts. Accordingly, from the ART EDB of Adama Hospital 18 predictor attribute and one class attribute with 10690 records of adult ART clients were selected based on domain expert suggestion and data mining goals. The selected attribute and there possible value presented following table 1 below:

Table 1: predictor attributes description by their type and possible value Selected ART Clients Records at Adama referral Hospital ART Clinic Sept 11/2005- April 24/2012

NO	Attributes	Data Type	Description	Possible values	Remark
1	**sex**	Nominal	Gender of a patient	Male or Female	
2	**Referral**	String	Facilities which the client referred to the ART clinic	From Adama Hospital or out of Adama Hospital	Aggregated
3	**Family Planning**	nominal	Whether the client uses any type of family planning methods	FALSE or TRUE	
4	**Eligible Reason**	Text	Criteria's for initiating of the ART regimen	CD4 count, clinically and both clinically and CD4 count	Aggregated
5	**OACD4**	numeric	Patient CD4 count at start of ART	Number	
6	**OA Status**	Ordinal	Functional Status of the patient	A(Ambulatory), B(Bedridden) or W(Working)	
7	**Age**	Numeric	Age of a patient at the start of ART	Number	
8	**OA Weight**	Numeric	Weight of a patient at the start of ART	Number	
9	**Current Regimen**	Character	ART regimen currently the client taking	1a,1b,1c,1d,1e,1f, 1gOTH,2a,2e	
10	**Has Family**	Nominal	The presence of children and other family member	FALSE OR TRUE	
11	**Adherence**	Nominal	Adherence status of the patient	Adherent or Non-Adherent	Aggregated
12	**Marital Status**	Nominal	Marital Status	Never Married, Married, Separated, Divorced or Widow	
13	**Educational Level**	String	Level of education	No Education, Primary, Secondary or Tertiary	
14	**Religion**	Nominal	Religion of the patient	Orthodox, Muslim, Protestant, Catholic or Other	
15	**Month on ART**	numeric	The number of months the client taking ART	Number	Aggregated
16	**Address**	String	Address of the Patient	In Adama town or out of Adama	Aggregated

				town	
17	**Occupation**	String	employment status of the patient	Non employed, civil servant, self employed, farmer, student	Aggregated
18	**OAWHO**	Ordinal	Patient WHO stage at start of ART	1,2,3,4	
19	**Outcome**	Nominal	The outcome of the client at time of data collection	DEAD,OA,TO,LO, DO	Aggregated

4.3.3 Data Transformation and Aggregation

In this step, data were merged, codified and transformed into forms appropriate for data mining. This task has implication in making the data useable and navigable for classification. In other word ,"the predictive power of data resides in transformation of the data, rather than in the raw data itself"(24). The Adama Hospital ART clinic EDB information (MS Access database) was exported into MS excel. The exported MS Excel includes the attributes from the registration and follow up table. The column contains 52 attribute and the row includes 19087 records. For the attributes, which duplicated by its nature the columns, were merged based and produce aggregated attribute. For example, attributes like ART started date and last appointment date merged by taking the difference between the two attributes value and produce month on ART attribute. For rows, which are, not relevant for this research eliminated accordingly based on the data set target criteria. The targeted data set discretized into categories and again the data set coded in to number form as shown in the annex IV below.

This exportation, aggregation, reduction and coding helps us to put the records in coma-separated form and for better classification performance. Then the coma delaminated formatted records were transformed into ARFF format(42). All the above tasks were to make the dataset easily navigable in to the WEKA data mining software. The ARFF format was used in the model building partially looks like as follow in figure 1:-

```
@RELATION DROPOUTART

@ATTRIBUTE sex {0, 1}
@ATTRIBUTE ReferralID {0, 1}
@ATTRIBUTE FamilyPlanningYN {0, 1}
@ATTRIBUTE EligibleReasonID {0, 1, 2}
@ATTRIBUTE OACD4 {0, 1, 2, 3}
@ATTRIBUTE OAStatus {0, 1, 2}
@ATTRIBUTE AgeInYears {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
@ATTRIBUTE OAWeight {0, 1, 2, 3, 4}
@ATTRIBUTE CurrentRegimen {0, 1, 2, 3, 4, 5, 6, 7}
@ATTRIBUTE HasFamily {0, 1}
@ATTRIBUTE Adherence {0, 1}
@ATTRIBUTE MaritalStatusID {0, 1, 2, 3, 4}
@ATTRIBUTE EducationalLevelID {0, 1, 2, 3}
@ATTRIBUTE ReligionID {0, 1, 2, 3, 4}
@ATTRIBUTE MonthonART {0, 1, 2, 3}
@ATTRIBUTE Address {0, 1}
@ATTRIBUTE Occupation {1, 2, 3, 4, 5}
@ATTRIBUTE OAWHO {1, 2, 3, 4}
@ATTRIBUTE Outcome {DEAD, OA, TO, DO, LO}

@DATA
0,1,0,0,0,1,5,1,0,1,1,4,0,1,0,1,4,4,DO
1,1,0,2,0,2,4,2,1,1,1,1,1,1,1,0,1,3,DO
1,1,0,0,0,1,7,3,0,1,1,1,1,1,1,1,4,3,DEAD
1,1,1,1,2,1,4,2,1,0,0,1,0,1,3,0,3,3,OA
1,0,0,0,0,2,3,3,0,1,1,3,2,1,1,1,2,2,DO
1,0,0,1,2,2,3,4,2,1,1,1,1,1,1,0,3,3,DO
0,1,0,0,0,1,2,2,0,1,1,3,1,2,0,1,3,3,DO
0,0,0,1,1,2,4,3,0,1,1,3,1,1,1,0,3,3,DO
1,0,0,1,0,1,3,3,0,0,1,0,3,1,2,1,3,3,DO
0,0,0,1,0,1,6,1,0,0,1,3,1,1,1,0,4,3,DEAD
0,0,0,0,1,0,2,1,0,1,1,3,2,1,0,1,4,4,DO
0,1,0,1,0,2,2,1,1,1,1,3,1,1,0,1,4,4,DO
1,1,1,1,1,2,1,3,2,0,0,0,2,1,2,1,4,3,DO
0,1,0,2,0,1,2,2,0,1,1,1,2,1,1,0,4,4,DO
```

Figure 1: Format of attributes label with sample data (ARFF)

5. Ethical considerations

Approvals to conduct the study were obtained from Institutional ethical Review Board of institution of Public health, University of Gondar. Official letters were taken from the University of Gondar to communicate and get permission from the leadership of Adama referral Hospital ART clinic. In order to maintain patients record confidentiality de-identified (a data set without name, unique ART number and MRN) electronic database were taken for this research purpose. During extraction of EDB data manager from the ART clinic were utilized. Therefore, ethical issue can easily understandable by those data collector and the principal investigator strongly addresses the issues of maintaining client's records confidentiality in training secessions.

6. RESULT

6.1. Socio-Demographic Characteristics of the ART Clients Record

From the existing 19,087 records of Adama referral Hospital ART clinic EDB including both pre-ART and ART 10,690 records were reviewed. The records accounts the total number of available ART client records. Female clients were 55.5% of the total client were registered in EDB and 44.5 % of the clients were male. Clients out of Adama town are more than clients live in Adama town by 29.2 %. Orthodox in religion clients record accounts 77.3 % the total records. Married client records were greater than never married client records by 46.3 %, As shown in the table 2 below:

Table 2: Socio-Demographic Characteristics of Selected ART Clients Records at Adama referral Hospital ART Clinic Sept 11/2005- April 24/2012

Variables		Number(N=10960)	Percent (%)
Sex	Female	5934	55.5
	Male	4756	44.5
	Total	10690	100
Address	Adama	4315	40.4
	Out of Adama	6375	69.6
	Total	10690	100
Marital Status	Divorced	1840	17.2
	Married	5109	47.7
	Never Married	160	1.4
	Separated	1910	17.8
	Widow	1671	15.6
	Total	10690	100
Education	No Education	2600	24.3

	Primary	4107	38.4
	Secondary	3224	30.1
	Tertiary	759	7.1
	Total	10690	100
Age	15-19	128	1.1
	20-24	978	9.1
	25-29	2284	21.3
	30-34	2265	21.1
	35-39	2115	19.7
	40-44	1242	11.6
	45-49	782	7.3
	50+	896	8.3
	Total	10690	100
Religion	Muslim	1296	12.1
	Orthodox	8267	77.3
	Protestant	1051	9.8
	Other	76	0.7
	Total	10690	100

The records from Adama referral Hospital ART clinic classified into five outcome variables. Accordingly among the records 350 (3.3%) were dead, 5427(50.8%) were on ART, 1855 (17.4%) were transfer out to other health facilities, 2784 (26.0%) were drop out from ART services and 274 (2.6%) were lost follow up from the ART clients record of the EDB were utilized. The ART clients on the selected record have mean age of 35 year. The descriptive statics of selected predictor variable and there cross tabulation with outcome variable shown on table 3 below.

Table3. Treatment and services outcomes of ART Clients Records at Adama Hospital ART Clinic Jan 11/2005- Apr 24/2012

	Out come										TOTAL	
	DEAD		DOPOUT		LOST		ON ART		TRANSFER OUT			
sex	N	%	N	%	N	%	N	%	N	%	N	%
M	185	3.9%	1397	29.3%	133	2.8%	2255	47.4%	787	16.5%	4756	44.5%
F	165	2.8%	1387	23.3%	141	2.4%	3173	53.5%	1068	18%	5934	55.5%
Functional Status												
B	106	10.4%	465	45.7%	12	1.2%	293	28.8%	142	13.9%	1018	9.5%
A	145	4.5%	1030	32%	80	2.5%	1368	42.5%	595	18.5%	3218	30.1%
W	99	1.5%	1289	19.9%	182	2.8%	3766	58.5%	1118	17.3%	6454	60.4%
Educational Level												
No Education	59	2.3%	785	30.2%	61	2.3%	1137	43.7%	558	21.5%	2600	24.3%
Primary	157	3.8%	1124	27.3%	111	2.7%	2020	49.2%	695	16.9%	4107	38.4%
Secondary	116	3.6%	745	23.1%	82	2.5%	1794	55.6%	487	15.1%	3224	30.4%
Tertiary	18	2.4%	130	17%	20	2.5%	477	62.8%	115	15.2%	759	7.1%
Address												
Adama	244	3.5%	1815	25.7%	166	2.3%	3630	51.5%	1195	16.9%	7050	65.9%
out of Adama	106	2.9%	969	26.6%	108	2.9%	1797	49.4%	660	18.1%	3640	34.1%
Occupation												
Jobless	42	3.8%	281	25.2%	25	2.2%	589	52.8%	178	16%	1115	10.4%
Civil servant	116	3.4%	830	23.9%	99	2.9%	1902	55%	514	14.9%	3461	32.4%
privet	25	2.2%	408	35.5%	40	3.5%	437	38.1%	238	20.7%	1148	10.7%
farmer	164	3.4%	1242	25.5%	108	2.2%	2443	50.3%	902	18.6%	4859	45.5%
student	3	2.8%	23	21.7%	2	0.9%	55	52.8%	23	21.7%	106	1%

6.2. **Model building and model evaluation**

For Model building in predicting the likely occurrence of dropout from ART and other outcomes of Adama Hospital ART clients, twelve experiments were done, where six of the experiments for constructing decision trees using J48 algorithm and the remaining six were using the neural network algorithm (MLP) to compare it with J48 algorithm, as shown in table 4 below**.**

Table 4 Test model parameters and data set type for model building on Treatment and services outcomes of ART Clients Records at Adama Hospital ART Clinic Jan 11/2005- Apr 24/2012

Experiments	Algorithm	Test model	Data set type
1	J48	Default (66% for training and 34% for testing)	Un balanced
2	J48	stratified 10-fold cross-validation	Un balanced
3	J48	stratified 10-fold cross-validation	Balanced in using SMOTE
4	J48	Default (66% for training and 34% for testing)	Balanced in using SMOTE
5	J48	Adjusted (90% for training and 10% for testing)	Un balanced
6	J48	Adjusted (90% for training and 10% for testing)	Balanced in using SMOTE
7	MLP	Default (66% for training and 34% for testing)	Un balanced
8	MLP	stratified 10-fold cross-validation	Un balanced
9	MLP	stratified 10-fold cross-validation	Balanced in using SMOTE
10	MLP	Default (66% for training and 34% for testing)	Balanced in using SMOTE
11	MLP	Adjusted (90% for training and 10% for testing)	Un balanced
12	MLP	Adjusted (90% for training and 10% for testing)	Balanced in using SMOTE

7.2.1 Model building using J48 algorithm

All the models from J48 algorithm were pruned tree at 0.25 confidence of pruning level for training and with minimum number of object per leave was two. In experiment one

10,690 instance (DEAD= 3.3%, OA= 50.8 %, TO= 17.4%, DO = 26.0%, LO = 2.6%) with 19 attributes were inputted in to the WEKA software in order to build decision tree model using J48 algorithm by using default test model parameters (training set 66% and test set 34%). The data set were un balanced. The result as shown in table 5 below has scored 73.7% accuracy.

The next step following model building using the default value of WEKA software Iterative model building and assessment were continue using readjusted the percentage of dataset for training and testing randomly and cross validation rule until, the principal investigator strongly believe that the best model(s) is found. These iterative procedures were done for both MLP and J48 algorithms. In this manner experiment five carried out by applying training set 90% and test set 10% data set inputted into J48 algorithm. The experiments were found 72.1% predicting accuracy in using the inputted 10,690 instances.

Table 5: Input Parameters and the Resulting J48 Decision Trees' Output on Treatment and services outcomes of ART Clients Records at Adama Hospital ART Clinic Jan 11/2005- Apr 24/2012

Experiment	No. of Attribute	Tree size	Number of leaves	Time taken for Experimentation	kappa	Area under ROC	Accuracy Score
1	19	838	635	0.67	0.5661	0.867	73.7%
2	19	838	635	0.8	0.5681	0.859	73.5%
3	19	7488	5686	6.46	0.9126	0.976	93.0%
4	19	7488	5686	7.29	0.9005	0.972	92.1%
5	19	838	635	0.87	0.5305	0.888	72.1%

6	19	7488	5686	7.1	0.9117	0.976	93.0%

In experiment three and six decision tree found from J48 algorithm in using balanced data set and using stratified 10-fold cross-validation. Moreover, re adjusted test model parameter (training set 90% and test set 10%) respectively.

Due to the reason that unbalance dataset for each class labels cause minority class classified incorrectly and showed low accuracy in all over prediction performance of the models, the dataset were balanced .These balancing of the dataset were done by applying Synthetic Minority Over-sampling Technique (SMOTE) algorithm analysis until the accuracy decline. This iterative process accounts 16 SMOTE experiments. From the analysis the data set accompanied 65,566 instances (DEAD= 17.1%, OA= 16.5 %, TO= 22.6%, DO = 17%, LO = 26.7%) with in 19 attributes. As a result, both experiment three and experiment six showed 93% accuracy.

From the twelve experiments, experiment three shows better performance for predicting dropout from ART (based on Kappa, time taken for experimentation and showing false alarm rate from confusion matrix). For this reason experiment three further evaluated using other performances evaluation mechanisms like recall and precision. As shown in table 6 weighted precision of the experiment three was 0.93.

Table 6: Precision and Recall Accuracy Measures for Model Built on Experiment 3 Using Default Parameter Values of J48 Algorithm on Treatment and services outcomes of ART Clients Records at Adama Hospital ART Clinic Jan 11/2005-Apr 24/2012.

	Precision	**Recall**
Dead	0.966	0.972
On ART	0.898	0.924
Transfer out	0.923	0.947
Dropout from ART	0.859	0.797

Lost follow up from ART	0.978	0.98
Weighted average	0.93	0.931

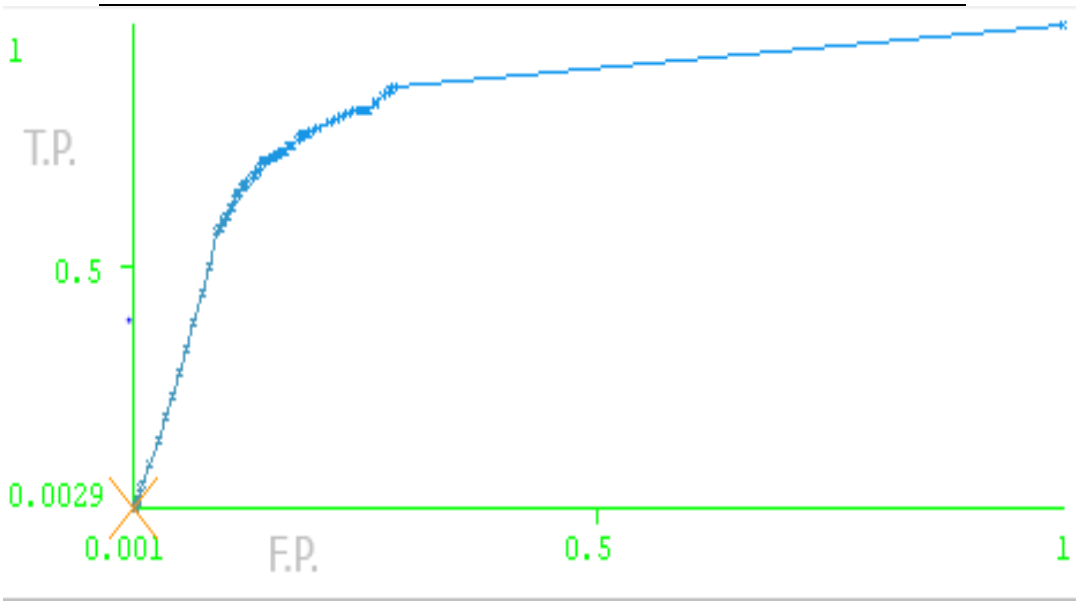

Figure: 2 The receiver operator curve from J48 algorithm from experiment three on Treatment and services outcomes of ART Clients Records at Adama Hospital ART Clinic Jan 11/2005- Apr 24/2012

On the other hand, the values of correctly classified records of experiment three depicted in using the confusion matrix as shown in table 7.

In The confusion matrix out of the total records provided to the J48 algorithms (65,567), 61,033 records (93.1%) were correctly classified into the intended outcome variables and among them 10,890 (97.2%) records were classified correctly in the class of dead. 10,854 (92.3%) records were classified correctly in the class of on ART. 14841 (94.7%) correctly classified into transfer out category and 11136 (79.7%) were classified correctly as dropout from ART class label in the J48 algorithm.

Table 7: confusion matrix for Model Built on Experiment 3 Using Default Parameter Values of J48 Algorithm on Treatment and services outcomes of ART Clients Records at Adama Hospital ART Clinic Jan 11/2005- Apr 24/2012 .

Actual	Predicted					Total	Accuracy rate
	Dead	On ART	Transfer out	Dropout from ART	Lost follow up from ART		
Dead	10890	0	105	203	2	11200	97.2%
On ART	1	10024	2	530	297	10854	92.3%
Transfer out	96	2	14059	677	7	14841	94.7%
Dropout from ART	287	853	1049	8870	77	11136	79.7%
Lost follow up from ART	3	287	12	44	17190	17536	98.0%
Total	11277	11166	15227	10324	17573	65567	93.1%

On the other hand, 105 (1%) records were incorrectly classified as Transfer out , 203(2%) records were incorrectly classified as Dropout from ART, 2(0%) records were incorrectly classified as Lost follow up from ART while actually they were supposed to be in the dead class. Similarly, 1(0%) records were incorrectly classified as Dead, 2 (0%) records were incorrectly classified as Transfer out, 530 records were incorrectly classified as Dropout from ART, 297 (2.7) records were incorrectly classified as Lost follow up from ART while actually they were supposed to be in the On ART class.

The experiment three conducted using SMOTE still has showed high number of incorrectly classified instance, especially for dropout class label. To overcome the problem related to unbalanced data that makes the minority class label predicted incorrectly, six re-sampling of instance with replacement on the dataset from the SMOTE was conducted and the accuracy increased into 95% from 93%.

In addition, the values of correctly classified records of experiment three using six re-sampling of instance with replacement and 10-fold cross validation depicted in using the confusion matrix as shown in table 8.

In The confusion matrix out of the total records provided to the J48 algorithms (65,567), 64,426 records (98.3%) were correctly classified into the intended outcome variables and among them 10,427 (99.6%) records were classified correctly in the class of dead. From instances in class of on ART 10,512 (97.9%) records were classified correctly in the class of on ART. From class of transferred out 15,325 (98.8%) correctly classified into transfer out category and 11,651 (94.6%) were classified correctly as dropout from ART class label in the J48 algorithm.

Table 8: confusion matrix for Model Built on Experiment 3 Using re-sampled dataset, 10-fold cross validation and J48 Algorithm on Treatment and services outcomes of ART Clients Records at Adama Hospital ART Clinic Jan 11/2005-Apr 24/2012.

Actual	Predicted					Total	Accuracy rate
	Dead	On ART	Transfer out	Dropout from ART	Lost follow up from ART		
Dead	10384	0	17	26	0	10427	99.6%
On ART	1	10291	1	138	81	10512	97.9%
Transfer out	33	0	15147	143	2	15325	98.8%
Dropout from ART	102	214	283	11024	28	11651	94.6%
Lost follow up from ART	0	61	5	6	17580	17652	99.6%
Total	10520	10566	15453	11337	17691	65567	98.3%

The decision tree were found from the J48 algorithm in experiment three is very large in size with5686 number of leaves and 7488 tree size, Which is too complex to generate rule. In order to minimize size of decision tree the principal investigator tries to modify the default value of minNumObj (minimum number of instances in a leaves) of WEKA software to 20, which were 2 in its default value. This modification was producing the decision tree with the number of 20 records at each leaves. The decision

tree size reduced in to 2047 number of leaves and 2640 tree size. The decision tree looks like partially in figure 3 below

```
J48 pruned tree
------------------

Adherence = 0
|   ReligionID = 0
|   |   OAWHO = 1: OA (86.0/64.0)
|   |   OAWHO = 2: OA (18.0/2.0)
|   |   |   OAWeight = 2: DO (76.0/5.0)
|   |   |   OAWeight = 3: OA (147.0)
|   |   |   OAWeight = 4: OL (184.0/65.0)
|   |   OAWHO = 3: OA (832.0/472.0)
|   |   OAWHO = 4: LO (287.0/225.0)
|   ReligionID = 1
|   |   FamilyPlanningYN = 0: LO (2888.0/1335.0)
|   |   FamilyPlanningYN = 1
|   |   |   ReferralID = 0: LO (11342.0/4759.0)
|   |   |   ReferralID = 1:
|   |   |   |   AgeInYears = 0: OA (6.0)
|   |   |   |   AgeInYears = 1: LO (301.0/203.0)
|   |   |   |   AgeInYears = 2: LO (704.0/551.0)
|   |   |   |   AgeInYears = 3: LO (674.0/526.0)
|   |   |   |   AgeInYears = 4: LO (560.0/487.0)
|   |   |   |   AgeInYears = 5: OA (311.0/23.0)
|   |   |   |   AgeInYears = 6: OA (146.0/16.0)
|   |   |   |   AgeInYears = 7: OA (42.0/21.0)
|   |   |   |   AgeInYears = 8: OA (47.0/12.0)
|   |   |   |   AgeInYears = 9: OA (14.0/9.0)
|   |   |   |   AgeInYears = 10: OA (12.0/7.0)
|   ReligionID = 2: OA (1110.0/188.0)
|   ReligionID = 4: OA (31.0/2.0)
Adherence = 1
|   MonthonART = 0
|   |   ReligionID = 0: DO (141.0/108.0)
|   |   ReligionID = 1
|   |   |   OAStatus = 0: DEAD (2484.0/863.0)
|   |   |   OAStatus = 1
|   |   |   |   CurrentRegimen = 0: DEAD (2991.0/2041.0)
|   |   |   |   CurrentRegimen = 1: TO (177.0/86.0)
|   |   |   |   CurrentRegimen = 2: DEAD (236.0/28.0)
|   |   |   |   CurrentRegimen = 3: DO (27.0/3.0)
|   |   |   |   CurrentRegimen = 7: TO (4.0)
```

Figure 1: Partial View of Decision Tree for Model Built on Experiment three Using Default Parameter Values of J48 Algorithm on Treatment and services outcomes of ART Clients Records at Adama Hospital ART Clinic Jan 11/2005- Apr 24/2012

Nevertheless, reducing the tree size has its own disadvantage, the accuracy of the prediction performance reduce from 98.3% to 90.4%. The complexity of the decision tree from the experiment three with default value of 2 minNumObj (minimum number of instances in a leave) out weighted by its accuracy of prediction. In addition, the result from the confusion matrix showed 57.2% of drop out records classified correctly the remaining 42.8% were classified incorrectly in to the classes of on ART, dead, lost follow up and transfer out, while actually they were supposed to be in the dropout class, as shown on table 9.

Table 9: Confusion matrix for Model Built on Experiment 3 Using adjusted Parameter Values (minNumObj) of J48 Algorithm on Treatment and services outcomes of ART Clients Records at Adama Hospital ART Clinic Jan 11/2005- Apr 24/2012.

Actual	Predicted					Total	Accuracy rate
	Dead	On ART	Transfer out	Dropout from ART	Lost follow up from ART		
Dead	10745	0	193	259	3	11200	95.9%
On ART	2	9590	0	536	726	10854	88.4%
Transfer out	351	1	13219	1239	30	14841	89.1%
Dropout from ART	799	1491	2293	6374	179	11136	57.2%
Lost follow up from ART	5	458	13	28	17032	17536	97.1%
Total	11277	11166	15227	10324	17573	59960	86.9%

Moreover, to get better accuracy and to minimize the decision tree complexity additional experiment were conducted by selecting important attribute using attribute sub set evaluator (CfsSubsetEval). Accordingly, eight attribute selected as best predictor sub set with merit of 0.44. The selected attributes were Referral ID, OA Status, Age In Years, Current Regimen, Adherence, Educational Level ID, Religion ID

and Address. The experiment conducted Based on the selected attribute J48 algorithm with balanced data set and 10-fold cross validation test parameter were found an accuracy of 74.4% , which is low in accuracy than the selected experiment three.

7.2.2 Model building using multilayer perceptron (MLP)

Experiments seven up to twelve were done in using MLP algorithm. In experiment seven the unbalanced data set were inputted into the MLP algorithm analyzed in using the default parameters of test model (training set 66% and test set 34%). The experiment seven were found 70.5% accuracy. Experiment eleven conducted with the inputted unbalanced data set by re adjusting the test model parameter (training set 90% and test set 10%) and 72.1% prediction accuracy were found as shown in table 10 below.

Table 10 : Input Parameters and the Resulting MLP Output on Treatment and services outcomes of ART Clients Records at Adama Hospital ART Clinic Jan 11/2005- Apr 24/2012

Experiment	Number of Attribute	Time taken for Experimentation	Kappa	Area under ROC	Accuracy Score
7	19	1326.99	0.5219	0.861	70.5%
8	19	260.92	0.5221	0.865	70.2%
9	19	1458	0.681	0.916	74.8%
10	19	2766.76	0.6774	0.916	74.5%
11	19	2720.78	0.5091	0.888	70.2%
12	19	1266.54	0.6804	0.914	74.8%

On the other hand, in order to find high prediction performance rate from the MLP algorithm the principal investigator tries to balance the data set in using SMOTE

algorithm. From the SMOTE with eight iterative SMOTE experimentation 21,889 instance (DEAD= 12.8%, OA= 24.8 %, TO= 16.9%, DO = 25.4%, LO = 20.0%) were found until the memory of the computer handle the process result. Accordingly, the experiment nine and twelve conducted by inputting the balanced data set in using stratified 10-fold cross validation and re adjusted (training set 90% and test set 10%) test model parameter respectively and both experiments were showed prediction accuracy of 74.8%.

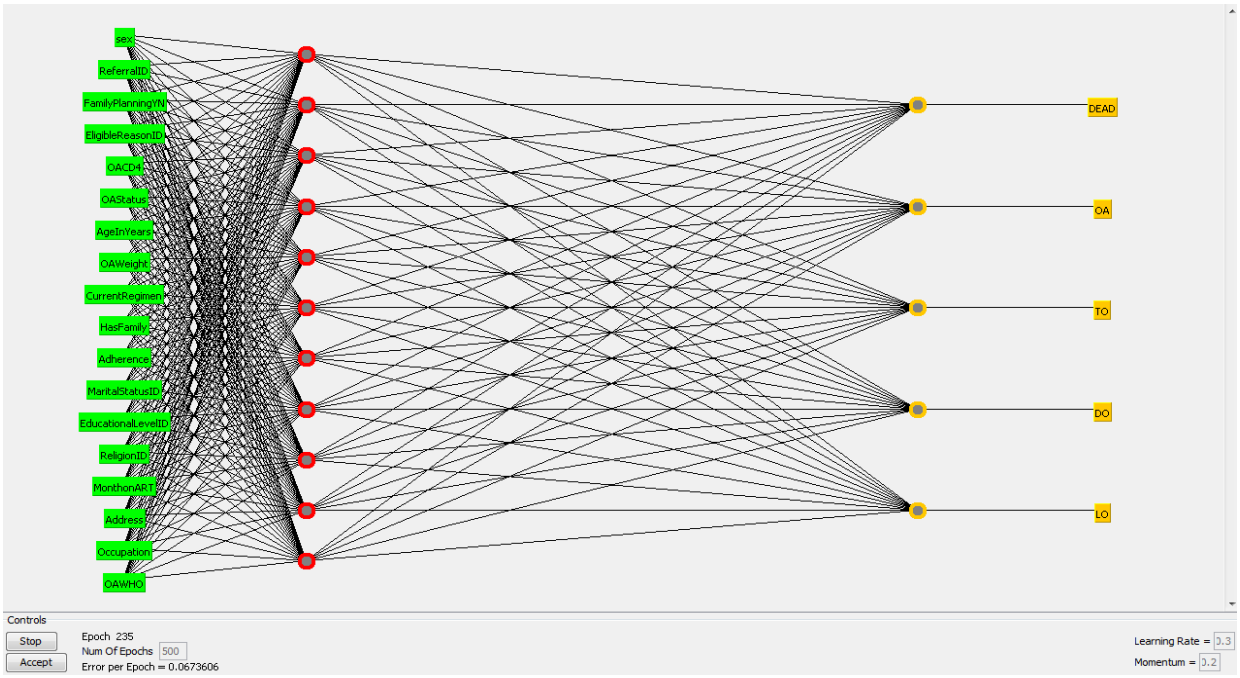

Figure 3: Feed forward neural network from the output of multilayer perceptron from experiment seven with the default value of test model on Treatment and services outcomes of ART Clients Records at Adama Hospital ART Clinic Jan 11/2005- Apr 24/2012.

Among the models build in using MLP experiment nine by inputting balanced dataset perform best prediction accuracy, the experiment eleven detail precision and recall depict in the table eleven below. Accordingly, weighted precision of experiment eleven was 0.74.

Table 11: Precision and Recall Accuracy Measures for Model Built on Experiment seven Using Default Parameter Values of MLP Algorithm on Treatment and services outcomes of ART Clients Records at Adama Hospital ART Clinic Jan 11/2005- Apr 24/2012.

	Precision	Recall
Dead	0.813	0.815
On ART	0.799	0.872
Transfer out	0.633	0.61
Dropout from ART	0.619	0.562
Lost follow up from ART	0.878	0.909
Weighted average	0.743	0.748

In addition, the values of correctly classified records of experiment nine (best performing experiment among MLP experiments) depicted in using the confusion matrix as shown in 12 table

In the confusion matrix out of the total records provided to the MLP algorithms, (21,889) 16,383 records (74.8%) were correctly classified into the intended outcome variables and among them 2281 (81.5%) records were classified correctly in the class of dead. From the imputed instances 5427 (87.2%) records were classified correctly in the class of on ART, and 3710(61.0%) correctly classified into transfer out category and 5568 (56.2%) were classified correctly as dropout from ART class label in the MLP algorithm.

Table 12: Confusion matrix for Model Built on Experiment nine Using Default Parameter Values of MLP Algorithm on Treatment and services outcomes of ART Clients Records at Adama Hospital ART Clinic Jan 11/2005- Apr 24/2012.

Actual	Predicted					Total	Accuracy rate

	Dead	On ART	Transfer out	Dropout from ART	Lost follow up from ART		
Dead	2281	0	160	353	6	2800	81.5%
On ART	3	4730	6	278	410	5427	87.2%
Transfer out	167	6	2262	1255	20	3710	61.0%
Dropout from ART	353	848	1120	3127	120	5568	56.2%
Lost follow up from ART	1	335	26	39	3983	4384	91.0%
Total	2805	5919	3574	5052	4539	21889	74.8%

On the other hand, 160 (5.7%) records were incorrectly classified as Transfer out, 353 (12.6%) records were incorrectly classified as Dropout from ART, six (0.2%)records were incorrectly classified as Lost follow up from ART while actually they were supposed to be in the dead class. Similarly, three (0%) records were incorrectly classified as Dead, six (0.1%) records were incorrectly classified as Transfer out, 278 (5.1%) records were incorrectly classified as Dropout from ART, 410 (7.6%) records were incorrectly classified as Lost follow up from ART while actually they were supposed to be in the On ART class.

6.3. Rules generated from the decision tree of experiment three

The output from J48 algorithm of experiment three can be putted in the form of if then rule. This is easy to understand by any person who can read it in the ART clinic. The rule goes through the decision tree and at the end of each leaves one rule can be produce with the possible case can be classified correctly and incorrectly. The predictor attribute were putted in rank manner for producing the rule. The rank is based on the calculated information gain as shown table 13.

Table 13: predictor attributes rank based on their information gain from experiment three on Treatment and services outcomes of ART Clients Records at Adama Hospital ART Clinic Jan 11/2005- Apr 24/2012.

Rank	Attribute	Information gain	Rank	Attribute	Information gain	Rank	Attribute	Information gain
1	Adherence	0.722	**4**	Current Regimen	0.1333	**7**	Religion	0.0754
2	Month on ART	0.3142	**5**	OA Status	0.0946	**8**	OACD4	0.0746
3	Family Planning	0.2255	**6**	OA Weight	0.086	**9**	OAWHO	0.0647

Some rule were generated from experiment three listed as follow

1. If Adherence client is **good** AND Month on ART <6 month AND client Religion is protestant AND client referred with in Adama Hospital from other than ART clinics:

   THEN will **DROUPOUT**

- The domain experts expectation was “if the adherence of client is good for the ART client will stay on ART services.”

2 If Adherence of the client is good AND Month on ART >2 years AND utilize Family Planning methods AND working in private organization AND the client religion is Muslim AND functional statues is **bedridden** AND **living in Adama town**:

THEN the client will **TRANSFER OUT** to other health facility

- The domain experts believed that “if a client functional status is bedridden mean spent most of the day time on his/her bed , and the client residencies is in Adama town the client will follow his treatment in Adama Hospital because the functional status reflects the client is under serious illness or disabilities. So, such kind of clients prefer Hospitals nearby their permanent residence ”

3 If Adherence of the client is **good** AND Month on ART <6 month AND client Religion is orthodox AND client **functiona**l **status is working** AND client age is between 20-24 years AND CD4 count is < 100ML/DL AND Weight >60kg AND:

THEN the client will be **DEAD**

- The domain experts believed that "if a client adherence is good and his/her functional status is working mean active in his day to day activity, the client will be on ART because the functional status is the reflection of treatment successes "

4 If Adherence of client is **poor** AND Religion of client is Muslim AND WHO clinical stage is one AND Eligible Reason for ART is clinically AND client functional status is working:

THEN client will be **ON ART**

- The domain expert's expectation was "if the adherence of client is poor for the ART client will dropout or lost from the ART services."

he above rules indicate the possible outcome of ART client with the value of predictor attribute. In addition, it is possible to draw more rules from the decision tree were putted in using J48 algorithm.

6.4. **Comparison of Decision Tree from J48 and Neural Networks from MLP**

Among the objectives of this study was to compare the performance of J48 algorithm and multilayer perceptron algorithms and to select the one, which performs the best. For this reason the principal investigator compare the decision tree from experiment one, two and five were compared with the neural network from the experiment seven, eight and eleven the fact that in both group the inputted data set (10690 records and unbalanced) and test parameter were similar.

Accordingly, experiment one from J48 and experiment seven from MLP analyses in using the default parameters of test model(training set 66% and test set 34%), the accuracy and AUR were73.7%, 0.859 and 70.5%, 0.861 respectively in which, J48

algorithm showed better accuracy. However, the difference was not significant at 0.05 significance level in paired t-test. In addition, the time taken for building the model decision tree from experiment one in using J48 algorithm very quick than the experiment seven in using MLP.

In addition, the accuracy from J48 algorithm with unbalanced dataset and training set 66% and test set 34% split was greater than the experiment 15 conducted RF algorithm with unbalanced dataset and 25 random tree inputting.

Based on the experiments conducted 10-fold cross validation and unbalanced data set for both algorithm J48 and MLP, in experiment two and eight, J48 showed better accuracy, which is 73.5%, than MLP algorithm 70.2%. The difference was significant at 0.05 significance level in two-tailed corrected paired t-test. In addition, the time take for model building in using MLP was higher than J48 algorithm.

On the other hand, experiments conducted training set 90% and test set 10%, and un balanced data set for both algorithm J48 and MLP, in experiment five and eleven, J48 showed better accuracy, which is 72.1%, than MLP algorithm 70.2%. However, the difference was not significant at 0.05 significance level in two-tailed corrected paired t-test.

Moreover, the accuracy of 83.1% were found from the combination of the two algorithm by multi scheme boosting mechanism. This is a better accuracy than both algorithms found individually using un balanced data set. The difference was significant at 0.05 significance level in two-tailed corrected paired t-test.

Even if, the result found by inputting the balanced data set from SMOTE analysis were difficult to make comparison the fact that the inputted instance were not similar in number, reason that MLP could not processed equal instance as J48 algorithm due to memory congestion. However, the comparison was done using experimenter user interface panel. The result showed that by inputting equal balanced data set (65,567) and 10 fold cross validation test parameter accuracy of 93% and 87.5% for both J48 and MLP; respectively. In addition, J48 algorithm performed better prediction accuracy than MLP at 0.05 significance level in two-tailed paired t-test.

Hence, the Adama referral Hospital ART clinic could employ the results of this experiment as an input for the decision making process on dealing outcomes of ART clients especially on early drop out from ART prevention.

6.5. **Evaluate results and Deployment**

This step includes - Assessment on data mining results with respect to business success criteria, Reviewing the process, Determine next step, Plan for deployment, Plan for monitoring and maintenance and Produce final report. In other word, mathematical model for prediction of drop out from ART will integrated with the clinic day to day activities(24). However, for deployment successes the result of the study will be presented to University of Gondar, Institution of Public Health. Also forwarded to Adama referral hospital, Oromiya Regional Health Bureau for deployment of the discovered hidden knowledge and distributed accordingly for governmental organizations and non-governmental organizations interested in the finding of subject matter. An attempt will be made to present the findings in different conferences and workshops and will be sent to publication on scientific journals.

7. DISCUSSION

7.1. Experiments done in using the classification algorithm

The experiments done by the principal investigator showed better classification accuracy in the experiment three among experiment done using J48 algorithm with 93.0% prediction accuracy. On the other hand, among experiments done by using MLP algorithm experiment nine produce high prediction accuracy 74.8 %. From the above scenario we can look that balanced data set has found higher prediction accuracy. The default testing model parameters stratified 10-fold cross-validation and the adjusted test model parameter (training set 90% and test set 10%) doesn't showed any difference in the accuracy in both algorithms.

The difficult issue in this research is in finding similar study for comparison. Moreover, the study conducted on ART in using data mining application were done by inputting different data set structure and number of attributes. Even though similar study not found for comparison but there were related study conducted on to predict Patient's response to ART in Rome, Italy(35). The investigators use a combination of expert rules, logistic regression and non-linear machine learning. In this investigation logistic regression found that AUC 0.76 and accuracy of 75.63% and RF found that AUC of 0.77 and accuracy of 76.16%. the above algorithms scoreless accuracy than the selected experiment three which is 93.0% prediction accuracy and0.9126 AUC (35). The selected experiment three has also better accuracy than study conducted in Stanford university on to predict infrequently occurring HIV mutation given that history of ART, from those approaches classification tree performs 56-67% of AUC measurements (28).

Another ,Study conducted in south Africa(37) on prediction of CD4 count change using support vector machine classification model has found an accuracy of 83% prediction which is less accuracy than the selected experiment in this study.

The possible explanation for the discrepancy between performance accuracy might be due to the difference in data base structure, number and type of attributes, number of records, number of class label and predictive algorithm.

7.2. **Predictor attributes in the experiment 3 Model**

In the selected experiment three decision tree shows adherence of client to the ART regimen is selected as the first predictor variable (inf. Gain=0.722). This result lays in accordance to domain experts strong suggestions to predict dropout and other outcome variables. Still to predict dropout the adherence by itself cannot determine the dropout from ART. So, for client non adherent to the ART religion (inf. Gain=0.0754) were used as splitting variable, in which for the clients religion specified as Muslim, orthodox and protestant further predicting variable is needed which is family planning utilization(inf. Gain=0.2255). If the religion category is Muslim and the client were utilizing family planning the client will be on ART. Similarly, their religion is catholic and other none listed in the above they will be on ART. The possible explanation by the domain experts was the number of records was few and most of the catholic religion live in the urban setting so they will have better information on the importance of staying in the ART services.

On other hand, for clients adherent to ART the successor predicting variable were month on ART. If month on ART is greater than 12 months, religion is protestant, family planning is not utilized and the CD4 count is less than 100cell/dl will drop out. The domain experts provide explanation as the months on ART determines the pill burden and burn out from consecutive follow up procedure, in the reverse determines the probability of dropout along with the third predictor variable religion fourth family planning and CD4 count.

Among 18 predictor attributes, attributes which has high correlation with class attributes, low inter correlation each other and high predictive value were eight in number (Referral ID, OA Status, Age In Years, Current Regimen, Adherence, Educational Level ID, Religion ID and Address). The subset was acceptable by the domain experts.

Accordingly, the domain experts were explaining the entire decision tree with entire predictor variable in similar fashion. In their investigation, they found as the decision tree for predicting drop out from ART acceptable and can be implemented in the ART clinic of Adama referral Hospital.

Knowledge gained from the experiment three using J48 algorithm

According to all experiments done religion and family planning were second and third variables succeeding to adherence in predicting whether the ART clients dropout from ART. Which is interesting finding based on the domain experts comments and their expectations. For example in rule one in the section 6.3, if the client is non adherent and his religion is Muslim and not utilizing family planning , on 1a regimen of ART, his/her WHO stage is three and having a body weight of less than 50kg, he/she will dropout from ART by 90.4% accuracy. In other words, family planning none utilize clients with other predictor variable will drop out by 90.4% accuracy. On the other hand, in rule 11of section 6.3, if the client is none adherent and his/r religion Muslim but utilizes family planning methods he or she will be on ART services by 90.7% accuracy of prediction.

In addition, religion shows great impact in the prediction of dropout from ART. For example in rule 12 and 13, even if the client non adherent if his religion is catholic or other than Muslim, orthodox or protestant he/she will be on ART by 100% and 92.9% accuracy of prediction, respectively.

Moreover the decision tree strengthen universally accepted impact of adherence of client to the services is very important to predict the likely occurrences of dropout from ART and other outcome variables by putting as the first predictor variable .

8. LIMITATIONS AND STRENGTHS

The limitations of this study mainly concentrate on not including other important predictor attribute value from manual client records. Due to the data collected from secondary source the data surfers from multiple missing values. For this reason, some important attributes were excluded from the analysis, such as anti tuberculosis treatments, cotrimoxazol preventive therapy and past ART .

Strength of the study lies on all records of ART client obtained from the electronic database was included in the analysis. This produces high prediction accuracy and avoids the over fitting problem. Adama Hospital data manager, for the integrated data consistency and accuracy of the attributes range value, verified the preprocessed data.

9. CONCLUSION

The results of the experiments carried out in this research using decision tree and multilayer perceptron have revealed that the technique of data mining is applicable in the process of predicting outcomes of ART services.

- Data collection, selection and cleaning were major tasks, which took most of the experimental time of the research. This is due to higher volume of the EDB and presence of multiple missing value, outliers and noise data.
- Among twelve experiments conducted, decision tree in using J48 algorithm in using balanced data set can able to predict the likely occurrence of drop out from ART among the new clients of Adama hospital ART clinic.
- MLP algorithm showed very slow and time consuming model building process. Moreover, the output from MLP is much inconvenient for interpretation and further utilization.
- According to the conducted experiments the decision tree approach were more applicable and appropriate to the problem domain since it provides additional features such as simple and easily understandable rules that can be used by non-technical health care professionals as well as health care planners and policy makers.
- Referral ID, OA Status, Age In Years, Current Regimen, Adherence, Educational Level ID, Religion ID and Address were the most significant predictor attribute among the imputed predictor variables.

In general, encouraging results were obtained by employing both MLP and decision tree approaches and showed good accuracy and performance in predicting outcomes OF ART variables in the Adama Hospital ART clinic clients records.

10. RECOMMENDATIONS

.

In the processes of this data mining application on ART client records of Adama Hospital, data mining application shows promising performance to predict the outcome variable including dropout from ART. Based on the findings from the experimentation the following recommendations can be forwarded as future research direction for researchers and to health care providers:

For health institution

- For Adama Hospital management, Integration of the Adama Hospital EDB with the decision tree result of experiment three using J48 algorithm could be more important to find intelligent database for preventing early dropout from ART services.
- For ART clinic staffs (ART nurses, physicians and counselors), health planer, and police makers, efforts of preventing dropout from ART in Adama Hospital ART clinic could be on the adherences of the clients. In addition' the counseling session could incorporate suitable way for addressing issues of ART in views of religious context and values.
- Data manager and data clerks of Adama Hospital could give emphasis for the electronic database of Adama referral Hospital ART clinic, complete records value and standardization of records range of value.

For researcher

- The decision tree found from the J48 algorithms were very large in size, for this reason other investigators could try other classification algorithms to come up with small size decision tree and better prediction accuracy performance.
- Other investigators could try the experimentation by adding attributes from manual records of ART client.

11. References

1. R. Poynton M, M. McDanie A. Classification of smoking cessation status with a backpropagation neural network. Journal of Biomedical Informatics [Internet]. 2006;39(6). Available from: www.elsevier.com/locate/yjbin

2. Elena L, Hapsatou T, Lauren MU, Xavier A, A. David P. Cost-Effectiveness of Preventing Loss to Follow-up in HIV Treatment Programs: A Côte d'Ivoire Appraisal. PLoS Medicine [Internet]. 2009 [cited 2012 Mar 9];6(10). Available from: http://www.plosmedicine.org/article/infoFjournal.pmed.

3. Global HIV/AIDS response. Epidemic update and health sector progress towards Universal Access Progress Report. 2011;

4. Group GHWATW. Human Resources for Health Implications of Scaling Up For Universal Access to HIV/AIDS Prevention, Treatment, and Care: Ethiopia Rapid Situational Analysis. 2010;

5. Federal HIV/AIDS Prevention and Control Office | Federal Ministry of Health. Strategic Plan II For Intensifying Multisectoral HIV and AIDS Response in Ethiopia 2010/11 – 2014/15 [Internet]. 2010 [cited 2012 Mar 17]. Available from: www.etharc.org/resources/download/finish/33/517

6. Kiflie Y, OYB A, Ayalew, Muluneh T. Evaluation of HIV/AIDS clinical care quality: the case of a referral hospital in North West Ethiopia. Oxford Journals Medicine, Int Journal for Quality in Health Care. Oxford Journals Medicine, Int. Journal for Quality in Health Care. 2011;21(5):356–62.

7. Helmut K, Yibeltal A, Aynalem A, Mesfin SM, Damen HM. Utilization of antiretroviral treatment in Ethiopia between February and December 2006: spatial, temporal, and demographic patterns. International Journal of Health Geographics [Internet]. 2007 [cited 2012 Mar 17];6(45). Available from: http://www.ij-healthgeographics.com/content/6/1/45

8. Taye T. B, Anders J. Outcomes of Antiretroviral Treatment: A Comparison Between Hospitals and Health Centers in Ethiopia. Journal of the International Association of Physicians in AIDS Care (JIAPAC). 2010;9(5):318–24.

9. Beyene BK, Deribe K, Hailekiros F, Biadgilign S. Defaulters from antiretroviral treatment in Jimma university specialized hospital, south west Ethiopia. [Trop Med Int Health. 2008] - PubMed - NCBI. 2008;13(3):328–33.

10. Ministry of health. Guideline for implementation of antiretroviral therapy. 2005 [cited 2012 Mar 11]; Available from: http://search.babylon.com/?babsrc=HP_Prot

11. John C, Dennis R-D, Paul W, Atieno O, Joseph N, Omary M, et al. Monitoring Adherence and Defaulting for Antiretroviral Therapy in 5 East African Countries:

An Urgent Need for Standards. Journal of the International Association of Physicians in AIDS Care (JIAPAC). 2007;7(4):193–9.

12. Helmut K, Paul j C, Mesfin Samuel M, Damen HM, Yibeltal A. Bibliography on HIV/AIDS in Ethiopia and Ethiopians in the Diaspora: The 2010 Update. Ethiop. J. Health Dev. [Internet]. 2010 [cited 2012 Mar 23];25(1). Available from: http://www.ajol.info/index.php/ejhd/article/view/69854/57934

13. Joseph Kwong-Leung Y, Solomon Chih-Cheng C, Kuo-Yang W, Anthony D H, Erik J S, Simon D M. True outcomes for patients on antiretroviral therapy who are “lost to follow-up” in Malawi. Bulletin of the World Health Organization. 2007;85(7):550–4.

14. Nuredin I. Evaluation hance of survival/death status among HIV positive people under anti retroviral therapy program: the case of Adama Hospital. Addis Ababa University Libraries Electronic Thesis and Dissertations: AAU-ETD [Internet]. 2007; Available from: http://etd.aau.edu.et

15. Avina S, Scott K. Access to Antiretroviral Therapy for Adults and Children with HIV Infectionmin Developing Countries:mHorizons Studies. Management Sciences for Health, Arlington, VA. 2010;125:305–15.

16. Solomon Chih-Cheng C, Joseph Kwong-Leung Y, Anthony David H, Chin-Nam B, Rose K-D. Increased mortality of male adults with AIDS related to poor compliance to antiretroviral therapy in Malawi. Tropical Medicine and International Health. 2008;13(4):513–9.

17. Steven Y. H, Anna J, Efraim D, Alfons B, Dawn P. Population-based Monitoring of HIV Drug Resistance in Namibia with Early Warning Indicators. J Acquir Immune Defic Syndr. 2010;55(4):27–31.

18. Pillai S, Good B, Richman D, Corbeil J. A new perspective on V3 phenotype prediction. AIDS Res. Hum. Retroviruses. 2003 Feb;19(2):145–9.

19. Sam H, Alan Y, Mark S. C. Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief. ScienceDirect NeuroImage. 2010;56(2):544–53.

20. Mu-Jung H. Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. Expert Systems with Applications. 2007;32:856–67.

21. Hao J, Wai-Ki C. Classifying DNA repair genes by kernel-based support vector machines. Biomedical Informatics. 7(5):257–63.

22. Tim TNH. Predicting HIV Status Using Neural Networks and Demographic Factors. University of the Witwatersrand, Johannesburg [Internet]. 2007 Feb 15

[cited 2012 Mar 27]; Available from: http://wiredspace.wits.ac.za/handle/10539/2010

23. Dolce G. et al. Clinical signs and early prognosis in vegetative state: A decisional tree, data-mining study. Brain Inj. 2008;22:617–23.

24. Anagaw S. Application of data mining technology to predict child mortality patterns: the case of Butajira rural health project (BRHP). Addis Ababa University Libraries Electronic Thesis and Dissertations: AAU-ETD!: [Internet]. Available from: http://etd.aau.edu.et/dspace/handle/123456789/1185

25. Witten H, Frank E. Data Mining Practical Machine Learning Tools and Technique [Internet]. Second. Morgan Kaufmann; 2005 [cited 2012 Jun 30]. Available from: http://www.cs.waikato.ac.nz/ml/weka

26. Daria P, Boris R, Robert S. M, Leslie L, Mark L. Classification of infectious diseases based on chemiluminescent signatures of phagocytes in whole blood. Artificial Intelligence in Medicine. 2011;52:153– 163.

27. Duen-Yian Y, Ching-Hsue C, Yen-Wen C. A predictive model for cerebrovascular disease using data mining. Expert Systems with Applications. 2011;38:8970–7.

28. Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonça A. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Res Notes. 2011 Aug 17;4:299.

29. Ray S. et al. A combined data mining approach for infrequent event : analyzing HIV mutation changes based on treatment history. [Comput Syst Bioinformatics Conf. 2006] - PubMed - NCBI. 2006;385–8.

30. Manaswini P, Ranjit KS. Multilayer Perceptron Network in HIV/AIDS Application. International Journal of Computer Applications in Engineering Sciences (IJCAES). 2011;1(1):41–8.

31. Sandri M, Zuccolotto P. Variable selection using random forests, Data analysis, classificaiton and the forward search. Springer. 2006;263–70.

32. Liaw A, Wiener M. Classification and Regression by random Forest. 2002 [cited 2012 Jun 30];2(3).

33. Breiman. L. Manual on setting up, using, and understanding random forests. 2002 [cited 2012 Jun 30]; Available from: http://oz.berkeley.edu/users/breiman/ Using_random_forests_V3.1.pdf

34. Vararuk A. , Petrounias I. Kodogiannis V. Data mining techniques for HIV/AIDS data management in Thailand. Journal of Enterprise Information Management. 2008;21(1):52–70.

35. Tefera H. Application of data mining technology to identify significant patterns in census or survey in Ethiopia. [Internet]. Addis Ababa University Libraries Electronic Thesis and Dissertations: AAU-ETD!: Available from: http://etd.aau.edu.et/dspace/handle/123456789/1013

36. Jinn-Yi Y, Chuan-Wei T, Tai-Hsi W. Using data mining techniques to predict hospitalization of hemodialysis patients. Decision Support Systems. 2011;50:439–48.

37. Leke Betechuoh B.et al using inverse neural networks for HIV adaptive control. International Journal of Computational Intelligence Research (IJCIR). 2007;3(1):11–5.

38. Adeeba K, Annapuri K, Rosma MD, Sameem AK, Basir A. The Prediction of AIDS Survival: A Data Mining Approach. Proceedings of the 2nd WSEAS International Conference on Multivariate Analysis and its Application in Science and Engineering. 2009;48–53.

39. Prosperi MCF, Altmann A, Rosen-Zvi M, Aharoni E, Borgulya G, Bazso F, et al. Investigation of expert rule bases, logistic regression, and non-linear machine learning techniques for predicting response to antiretroviral treatment. Antivir. Ther. (Lond.). 2009;14(3):433–42.

40. Altmann A. Antiretroviral Therapy Optimisation without Genotype Resistance Testing: A Perspective on Treatment History Based Models. PLoS ONE [Internet]. 2010 [cited 2012 Apr 4];5(10). Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2966424

41. Yashik S, Maurice M. Support vector machines to forecast changes in CD4 count of HIV-1 positive patients. Scientific Research and Essays. 2010;5(17):2384–90.

42. Leul W. The application of data mining in crime prevention: the case of Oromia police commission. Addis Ababa University Libraries Electronic Thesis and Dissertations: AAU-ETD! [Internet]. Available from: http://etd.aau.edu.et/dspace/handle/123456789/1019

43. Remco R B, et al. WEKA Manual for Version 3-7-0. 2009 [cited 2012 Mar 23]; Available from: https://svn.scms.waikato.ac.nz/svn/weka/tags/dev-3-7-0

44. Breault JL, Goodall CR, Fos PJ. Data mining a diabetic data warehouse. Artificial Intelligence in Medicine. 2002;26(1):37–54.

45. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explore. Newsl. 2009 Nov;11(1):10–8.

46. Larose DT. Data mining methods and models. John Wiley and Sons; 2006.

47. Pete C et al. CRISP-DM 1.0 Step-by-step data mining guide [Internet]. 2000 [cited 2012 Mar 23]. Available from: www.the-modeling-agency.com/crisp-dm.pdf

12. Annexes

Annex I. Information Sheet

Title of the Research Project: Application of Data mining on Prediction of drop out from anti-retroviral therapy among HIV/AIDS patients of Adama referral hospital, Ethiopia, 2012

Name of Principal Investigator: Nebiyou Wendwessen (BSc)

Name of the Organization: University of Gondar, College of Medicine and Health Sciences.

Introduction: This information sheet is prepared with purpose of giving an explanation the research project for Oromiya regional health bureau and the leader ship of Adama referral hospital ART clinic. In this information sheet, the concerned bodies stated in the above will became clear the procedures of the research project and kindly expected to give permission to conduct the research.

Purpose of the Research Project: To predict drop out from ART before it occurs and tackle significant predictor variables during the course of ART follow up in the Adama ART clinic, by having interesting decision rule that will be develop in using data mining techniques.

Procedure: to undertake this all adults age greater than 15 ever started ART in Adama hospital ART clinic and registered in EDB will taken as the source population and all the records of the clients will be reviewed.

Risk and /or Discomfort: since the research project relays on document review, clients of Adama referral hospital ART clinic will not suffer from any kind of risks and discomforts due to their document review.

Benefits: for the clients whose record will be reviewed in this study project there is no direct benefit that they may acquire during the courses of the investigation. However, indirectly the study will provide the important knowledge that putted in the database in the decision rule format and reduce the risk of drop out from ART. In other hand, it will provide new areas of solution for health care planers, policy makers in advance and steak holders working in PLHIV care and support.

Confidentiality: In multiple ways of data, collection of this study the confidentiality of each client will be kept by having de- identified EDB records. Data collectors will be selected among ART clinic services providers that include data clerks and ART data manager. Because, using ART clinic service provider limits further exposure of the client's record to other individuals

Person to contact: This research project will be reviewed and approved by the ethical committee of the University of Gondar. If you want to have more information, you can contact the committee through the address below. If you have any question study under taken you can contact any of the with ethical committee through following addresses of the principal investigator and the research advisors.

.

1. Mr. Nebiyou Wendwessen
   Tel: +251-912-20-25-18 / e-mail: nerwtmk@gmail.com

2. Dr. Berihun Megebiaw (MD, MPH), University of Gondar, college of medicine and health sciences, institute of public health: Advisor
   Tel: +251-912-12-71-73 / e-mail: beredomegaeiaw@gmail.com

3. Mr. Bikes Destaw (BSC, MPH) University of Gondar, college of medicine and health sciences, institute of public health: Advisor
   Tel: +251-910-87-55-42 / e-mail: bikedes@yahoo.com

Eyyyama qo’anna kaanaf waraqaa odefanno

Mata-duree qo’anna kanaa : data mining fayyadamuudhaan nammota ART irra gaddhiisanii deeman tilmaaman

Maqaa qoraata: Naabiyyu waandassan (Bsc)

Maqaa dhabattichaa: yuunivaarsitii goondar

Seensaa : waraaqaan odefanno kun kan qopha’ee birrooo egumsa fayyaa nannoo oromiyaa fi hoospitaala refeeralii adamaa kutaa bulchiinsa “ART” wa’ee qo’anna kanaa ibsuudhaafi. Haala waraaqa odefannoo kanaatiin qamootnii armaan olitti ibsamaan hundii haala adeemsa qo’anna kanaa ergaa hubaatani booda qo’annan kun akka gaggefamuuf eyyama akka kennan kaabajaan ni gaaafatamaa.

Faayidaa qo’anno kanaa: Data mining fayyadamuudhaan namoota ART gaadhisaani bahaaniif tilmaama gochuuf.

Adeemsa : qo’anno kanaa gaggesuuf namoota umrii isaani 15 ta’eefi ART calqaaban hoospitaala referaali adamaatti kiliniiika ART tti fi EDB galma’aan namoota qo’anno kanaa fi ragaan kadhimamatoota hundii ni sakkata’ama.

Midhaama: qo’anno kun ragaa dhukkubsatoota sakkata’uu irratti waan hunda’uuf midhaan dhukkubsatoota irratti qopha’uu hin jiruu.

Faayidaa: dhukkubsattonni qoraanoo kanaa qo’annicha irratti hirmaachuu isaanitti faayidaan kallatiin argaatan hin jiruu. Haata’uu malee qo’annon kun cinaagalaan beekumsa gahaa murtoo murtessudhaafi namoota ART gadhiisanii deeman tilmaamuuf nifayyada. Haala biraatiin poolisii baasudhaa fi qaamoota addada adda PLHIV qarqaaraniif ni gaargara.

Iccitii : iccitiin dhukkubsatoota kaan egaamu qo’anna kanaaf adda baasa raga EDB tiin fayyadamuu. Sassabdonni ragaa kan filaataman namoota kutaa ART hojjetaan kesaatti.

Namaa argachuu dandessan

Qo’annan kun kaan ragga’uu yuunivaarsitii goondaariin. Yoo gaffi ykn odefanno dabaalata baarbaadan namoota maqaan isaani armaan gadditti tarrefamee gaafachuu dandessan.

Naabiyyu waandassan

Lakk. Bilbila +251-912-20-25-18/ e-mail: nerwtmk@gmail.com

Dr. baarihuun megabaawu

Lakk. Bilbila: +251-912-12-71-73 / e-mail: beredomegaeiaw@gmail.com

Obbo biikis daastaaw

Lakk. Bilbila:: +251-910-87-55-42 / e-mail: bikedes@yahoo.com

Emayilii

Annex II Consent form

Hello! My name is Nebiyou Wendwessen, student of the institute of Public Health in the University of Gondar and conducting a research for the partial fulfillment of second degree on “Application of data mining in prediction of dropout from anti-retroviral therapy among HIV/AIDS patients of Adama referral Hospital, Ethiopia”.

The objective the aim of this study is to predict dropouts from antiretroviral therapy at Adama referral Hospital, Ethiopia using application of data mining..

Your Hospital ART clinic were selected because your hospital had enough records on ART clients that is suitable for Data mining application and with the hope that you will cooperate with me. I kindly requesting you to give me the de- identified electronic data base. I assure that all information gathered during the course of the study will be kept completely confidential. All the records from electronic data base of ART clinic that you are going to deliver to me will be de-identified. Only the principal investigator will have access to the records.

I the undersigned as the manager of the hospital have been well informed the objective of the study entitled “Application of data mining in prediction of dropout from anti-retroviral therapy among HIV/AIDS patients of Adama referral Hospital, Ethiopia”. Having been well explained and informed of the intentions and benefits of the study, I voluntarily consent to deliver the de-identified electronic data base for the study.

Name…… Signature……

Annex III. CRISP-DM process adopted from CRISP-DM Step-by-step data mining guide

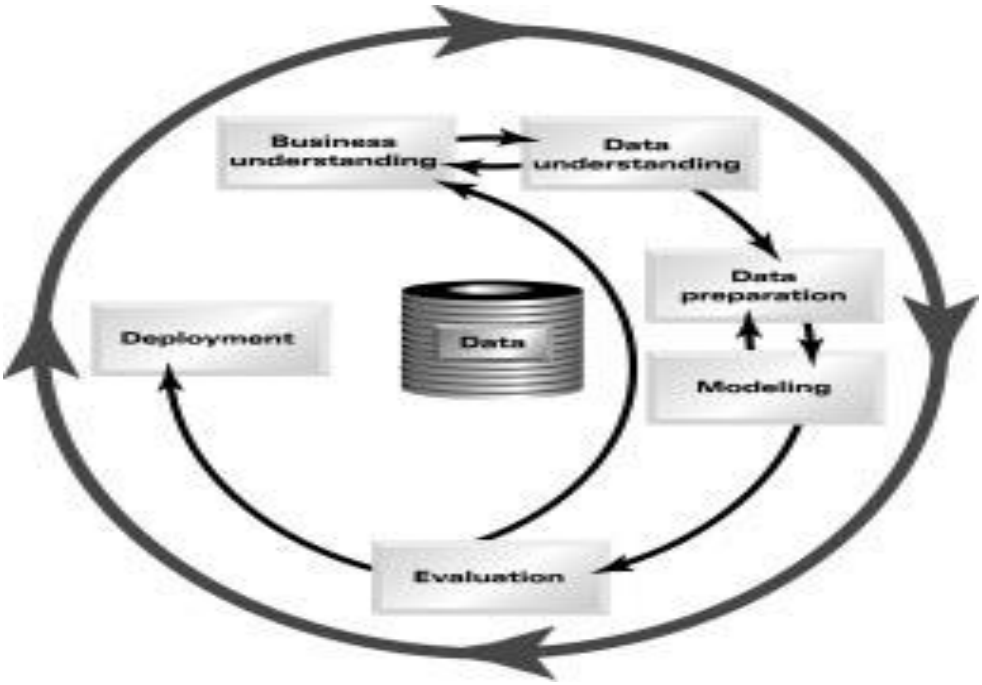

Figure 5: CRISP-DM processes

Annex IV Predictor variables data type, possible values and possible code during analysis on Treatment and services outcomes of ART Clients Records at Adama Hospital ART Clinic Jan 11/2005-Apr 24/2012

NO	Attributes	Data Type	Data Type	Possible values	codes
1	sex	Nominal	Text	Male or Female	M=1 F=0
2	Referral ID	String	Text	From Adama Hospital or out of Adama Hospital	From Adama Hospital=1 out of Adama Hospital=0
3	Family PlanningYN	nominal	Yes/ No	FALSE or TRUE	FALSE =0 TRUE=1
4	Eligible Reason ID	Text	Text	CD4 count, clinically and both clinically and CD4 count	clinically = 0 CD4 count= 1 both clinically and CD4 cou
5	OACD4	numeric	Number	Number	>100= 0 >200= 1 >350= 2 350+= 3
6	OAStatus	Ordinal	Text	A(Ambulatory), B(Bedridden) or W(Working)	B= 0 A= 1 W= 2

7	AgeInYears	Numeric	Number	Number	15-19= 0 20-24= 1 25-29= 2 30-34= 3 35-39= 4 40-44= 5 45-49= 6 50-54= 7 55-59= 8 60-64= 9 65+= 10
8	OA Weight	Numeric	Number	Number	>30= 0 >40= 1 >50= 2 >60= 3 60+= 4
9	Current Regimen	Character	String	1a(d4t, 3TC, NVP) 1b(d4t, 3TC, EFV) 1c(AZT, 3TC, NVP) 1d(AZT, 3TC, EFV) 1e(TDF, 3TC, NVP)	1a= 0 1b= 1 1c= 2 1d= 3 1e= 4

				1f(TDF, 3TC, EFV) 1g OTH(any other combination) 2a(ABC, ddI, NVP),2e	1f= 5 1gOTH= 6 2a and 2e= 7
10	Has Family	Nominal	Yes/ No	FALSE or TRUE	FALSE = 0 TRUE= 1
11	Adherence	Nominal	Text	Adherent or Non-Adherent	Adherent= 1 Non-Adherent= 0
12	Marital Status ID	Nominal	Text	Never Married, Married, Separated, Divorced or Widow	Never Married= 0 Married=1 Separated=2 Divorced= 3 Widow= 4
13	Educational Level ID	String	Text	No Education, Primary, Secondary or Tertiary	No Education= 0 Primary= 1 Secondary= 2 Tertiary= 3
14	Religion ID	Nominal	Text	Orthodox, Muslim, Protestant, Catholic or Other	Muslim= 0 Orthodox=1 Protestant= 2 Catholic= 3

					Other= 4
15	Month on ART	numeric	Number	Number	>6 month= 0 >1year= 1 >2year= 2 2years+= 3
16	Address	String	Text	In Adama town or out of Adama town	In Adama town= 1 out of Adama town= 0
17	Occupation	String	Text	Non employed, civil servant, self employed, farmer, student	Non employed= 1 civil servant=2 farmer=3 self employed=4 student=5
18	OAWHO	Ordinal	Number	1,2,3,4	Stage 1= 1 Stage 2= 2 Stage 3= 3 Stage 4= 4
19	Outcome	Nominal	Text	DEAD,OA,TO,LO,DO	Not codified

Annex V: J48 post manually pruned tree from experiment three on Treatment and services outcomes of ART Clients Records at Adama Hospital ART Clinic Jan 11/2005- Apr 24/2012

```
J48 pruned tree
------------------

Adherence = 0
|   ReligionID = 0
|   |   OAWHO = 1: OA (86.0/64.0)
|   |   OAWHO = 2: OA (18.0/2.0)
|   |   |   OAWeight = 2: DO (76.0/5.0)
|   |   |   OAWeight = 3: OA (147.0)
|   |   |   OAWeight = 4: OL (184.0/65.0)
|   |   OAWHO = 3: OA (832.0/472.0)
|   |   OAWHO = 4: LO (287.0/225.0)
|   ReligionID = 1
|   |   FamilyPlanningYN = 0: LO (2888.0/1335.0)
|   |   FamilyPlanningYN = 1
|   |   |   ReferralID = 0: LO (11342.0/4759.0)
|   |   |   ReferralID = 1:
|   |   |   |   AgeInYears = 0: OA (6.0)
|   |   |   |   AgeInYears = 1: LO (301.0/203.0)
|   |   |   |   AgeInYears = 2: LO (704.0/551.0)
|   |   |   |   AgeInYears = 3: LO (674.0/526.0)
|   |   |   |   AgeInYears = 4: LO (560.0/487.0)
|   |   |   |   AgeInYears = 5: OA (311.0/23.0)
|   |   |   |   AgeInYears = 6: OA (146.0/16.0)
|   |   |   |   AgeInYears = 7: OA (42.0/21.0)
|   |   |   |   AgeInYears = 8: OA (47.0/12.0)
|   |   |   |   AgeInYears = 9: OA (14.0/9.0)
|   |   |   |   AgeInYears = 10: OA (12.0/7.0)
|   ReligionID = 2: OA (1110.0/188.0)
|   ReligionID = 4: OA (31.0/2.0)
Adherence = 1
|   MonthonART = 0
|   |   ReligionID = 0: DO (141.0/108.0)
|   |   ReligionID = 1
|   |   |   OAStatus = 0: DEAD (2484.0/863.0)
|   |   |   OAStatus = 1
|   |   |   |   CurrentRegimen = 0: DEAD (2991.0/2041.0)
|   |   |   |   CurrentRegimen = 1: TO (177.0/86.0)
|   |   |   |   CurrentRegimen = 2: DEAD (236.0/28.0)
|   |   |   |   CurrentRegimen = 3: DO (27.0/3.0)
|   |   |   |   CurrentRegimen = 7: TO (4.0)
|   |   |   OAStatus = 2
|   |   |   |   AgeInYears = 1: TO (127.0/111.0)
|   |   |   |   AgeInYears = 2: DEAD (455.0/313.0)
|   |   |   |   AgeInYears = 3
|   |   |   |   |   EducationalLevelID = 0: TO (73.0/26.0)
|   |   |   |   |   EducationalLevelID = 1
|   |   |   |   |   |   OAWeight = 0: DO (20.0)
|   |   |   |   |   |   OAWeight = 1: DEAD (5.0)
|   |   |   |   |   |   OAWeight = 3: TO (28.0/16.0)
```

```
|   |   |   |   |   |   OAWeight = 4: TO (19.0/18)
|   |   |   |   |   EducationalLevelID = 2: DEAD (73.0/13.0)
|   |   |   |   AgeInYears = 4: DO (259.0/128.0)
|   |   |   |   AgeInYears = 5
|   |   |   |   |   EducationalLevelID = 0: DO (12.0)
|   |   |   |   |   EducationalLevelID = 1: DEAD (17.0/6.0)
|   |   |   |   |   EducationalLevelID = 2: TO (7.0)
|   |   |   |   AgeInYears = 6: DEAD (63.0/30)
|   |   |   |   AgeInYears = 7: TO (5.0/3.0)
|   |   |   |   AgeInYears = 8: DO (6.0)
|   |   |   |   AgeInYears = 9: TO (3.0)
|   |   ReligionID = 2: TO (99.0/44.0)
|   |   ReligionID = 3: DO (3.0)
|   MonthonART = 1
|   |   ReligionID = 0: DO (115.0/41.0)
|   |   ReligionID = 1
|   |   |   CurrentRegimen = 0
|   |   |   |   MaritalStatusID = 0: DEAD (188.0/119.0)
|   |   |   |   MaritalStatusID = 1
|   |   |   |   |   OAWeight = 0: DEAD (262.0/84.0)
|   |   |   |   |   OAWeight = 1
|   |   |   |   |   |   OAStatus = 0: DO (6.0)
|   |   |   |   |   |   OAStatus = 1: DEAD (17.0/6.0)
|   |   |   |   |   |   OAStatus = 2: TO (34.0/22.0)
|   |   |   |   |   OAWeight = 2: DEAD (336.0/135.0)
|   |   |   |   |   OAWeight = 3: TO (190.0/108.0)
|   |   |   |   |   OAWeight = 4: DO (43.0/29.0)
|   |   |   |   MaritalStatusID = 3: DO (212.0/30.0)
|   |   |   |   MaritalStatusID = 4: TO (90.0/32.0)
|   |   |   CurrentRegimen = 1: TO (91.0/58.0)
|   |   |   CurrentRegimen = 2: DO (48.0/34.0)
|   |   |   CurrentRegimen = 3: DO (13.0)
|   |   ReligionID = 2: TO (86.0/26.0)
|   MonthonART = 2
|   |   FamilyPlanningYN = 0
|   |   |   AgeInYears = 0: TO (2.0)
|   |   |   AgeInYears = 1: DEAD (115.0/102.0)
|   |   |   AgeInYears = 2
|   |   |   |   Occupation = 1: TO (12.0/7.0)
|   |   |   |   Occupation = 2: DEAD (139.0/60.0)
|   |   |   |   Occupation = 3: DO (25.0/16.0)
|   |   |   |   Occupation = 4: TO (159.0/21.0)
|   |   |   AgeInYears = 3
|   |   |   |   OACD4 = 0: DO (178.0/106.0)
|   |   |   |   OACD4 = 1: TO (67.0/5.0)
|   |   |   |   OACD4 = 2: TO (14.0/3.0)
|   |   |   AgeInYears = 4: DO (24.0/19.0)
|   |   |   AgeInYears = 5: DEAD (118.0/53.0)
|   |   |   AgeInYears = 6
|   |   |   |   sex = 0: TO (23.0/5.0)
|   |   |   |   sex = 1: DO (26.0/3.0)
|   |   |   AgeInYears = 7: DO (16.0/2.0)
|   |   |   AgeInYears = 8: DO (6.0)
|   |   FamilyPlanningYN = 1: TO (473.0/86.0)
```

```
|   MonthonART = 3
|   |   FamilyPlanningYN = 0
|   |   |   OAWHO = 1: TO (210.0/47.0)
|   |   |   OAWHO = 2: TO (680.0/312.0)
|   |   |   OAWHO = 3
|   |   |   |   EducationalLevelID = 0
|   |   |   |   |   OAWeight = 0: DO (44.0/24.0)
|   |   |   |   |   OAWeight = 1
|   |   |   |   |   |   Address = 0: DEAD (109.0/36.0)
|   |   |   |   |   |   Address = 1: DO (107.0/52.0)
|   |   |   |   |   OAWeight = 2
|   |   |   |   |   |   sex = 0
|   |   |   |   |   |   |   OAStatus = 0: TO (50.0)
|   |   |   |   |   |   |   OAStatus = 1
|   |   |   |   |   |   |   |   EligibleReasonID = 0: DO (81.0/57.0)
|   |   |   |   |   |   |   |   EligibleReasonID = 1: TO (144.0/27.0)
|   |   |   |   |   |   |   |   EligibleReasonID = 2: DO (84.0/41.0)
|   |   |   |   |   |   |   OAStatus = 2: TO (541.0/64.0)
|   |   |   |   |   |   sex = 1: DO (169.0/97.0)
|   |   |   |   |   OAWeight = 3: TO (464.0/200.0)
|   |   |   |   |   OAWeight = 4
|   |   |   |   |   |   ReferralID = 0: DEAD (20.0/17.0)
|   |   |   |   |   |   ReferralID = 1: TO (16.0)
|   |   |   |   EducationalLevelID = 1: TO (1729.0/1069.0)
|   |   |   |   EducationalLevelID = 2: DEAD (838.0/626.0)
|   |   |   |   EducationalLevelID = 3: TO (158.0/34.0)
|   |   |   OAWHO = 4: DO (808.0/773.0)
|   |   FamilyPlanningYN = 1
|   |   |   Occupation = 1: TO (786.0/252.0)
|   |   |   Occupation = 2: LO (2596.0/2445.0)
|   |   |   Occupation = 3: TO (1100.0/558.0)
|   |   |   Occupation = 4
|   |   |   |   EducationalLevelID = 0: TO (1157.0/294.0)
|   |   |   |   EducationalLevelID = 1
|   |   |   |   |   ReferralID = 0
|   |   |   |   |   |   OACD4 = 0
|   |   |   |   |   |   |   CurrentRegimen = 0: DEAD (296.0/219.0)
|   |   |   |   |   |   |   CurrentRegimen = 1: TO (49.0)
|   |   |   |   |   |   |   CurrentRegimen = 2: DO (47.0/29.0)
|   |   |   |   |   |   |   CurrentRegimen = 4: TO (45.0/19)
|   |   |   |   |   |   |   CurrentRegimen = 5: DEAD (3.0)
|   |   |   |   |   |   |   CurrentRegimen = 6: DO (3.0)
|   |   |   |   |   |   OACD4 = 1: TO (444.0/67.0)
|   |   |   |   |   |   OACD4 = 2: TO (174.0/27.0)
|   |   |   |   |   |   OACD4 = 3: TO (28.0)
|   |   |   |   |   ReferralID = 1: TO (752.0/82.0)
|   |   |   |   EducationalLevelID = 2: TO (803.0/325.0)
|   |   |   |   EducationalLevelID = 3: TO (11.0/10.0)
|   |   |   Occupation = 5: TO (98.0/9.0)
```

Declaration

I, the undersigned, senior MPH student declare that this thesis is my original work in partial fulfillment of the requirement for the degree of Master of Public Health in Health Informatics.

Name: Nebiyou Wendwessen

Signature: _____

Place of submission: istutet of public Health, College of Medicine and Health Sciences, University of Gondar.

Date of Submission: _____

This thesis work has been submitted for examination with our approval as university advisor(s).

Advisors

	Name	**Signature**	**Date**
1.	Dr. Berihun Megebiaw (MD,MPH)	_____	_____
2.	Mr. Bikes Destaw (BSc, MPH)	_____	_____

Assurance of Investigator

I, the undersigned, senior MPH student agree to accept responsibility for the scientific, ethical and technical conduct of the research project and for provision of required progress reports as pre terms and conditions of the research and publications office of the University of Gondar.

Name of the student: Nebiyou Wendwessen

Date: _____ Signature: _____

Approval of the advisor (s)

Advisors

Name	Signature	Date
1. Dr. Berihun Megebiaw (MD,MPH)	_____	_____
2. Mr. Bikes Destaw (BSc, MPH)	_____	_____